Optimal disclosure risk assessment

Published version:

DOI:10.1214/20-AOS1975

OPTIMAL DISCLOSURE RISK ASSESSMENT

By F. Camerlenghi[†,*], S. Favaro[‡,*], Z. Naulet[§] and F. Panero[¶]

University of Milano - Bicocca[†], *University of Torino*[‡], *University of Toronto*[§] *and University of Oxford*[¶]

Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. Consider a microdata sample of size n from a finite population of size $\bar{n} = n+\lambda n$, with $\lambda > 0$, such that each record contains two disjoint types of information: identifying categorical information and sensitive information. Any decision about releasing data is supported by the estimation of measures of disclosure risk, which are functionals of the number of sample records with a unique combination of values of identifying variables. The most common measure is arguably the number τ_1 of sample unique records that are population uniques. In this paper, we first study nonparametric estimation of τ_1 under the Poisson abundance model for sample records. We introduce a class of linear estimators of τ_1 that are simple, computationally efficient and scalable to massive datasets, and we give uniform theoretical guarantees for them. In particular, we show that they provably estimate τ_1 all of the way up to the sampling fraction $(\lambda+1)^{-1} \propto (\log n)^{-1}$, with vanishing normalized mean-square error (NMSE) for large n. We then establish a lower bound for the minimax NMSE for the estimation of τ_1, which allows us to show that: i) $(\lambda+1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction; ii) estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for large n. This is the main result of our paper, and it provides a precise answer to an open question about the feasibility of nonparametric estimation of τ_1 under the Poisson abundance model and for a sampling fraction $(\lambda + 1)^{-1} < 1/2$.

1. Introduction. Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. Any decision about release requires a careful assessment of the risk of disclosure, which is supported by the estimation of measures of disclosure risk (e.g., Willenborg and de Waal [27]). Consider a microdata sample $\boldsymbol{X}(n) = (X_1, \ldots, X_n)$ from a finite population $\boldsymbol{X}$ of size $\bar{n} > n$, and without loss of generality assume that each sample record X_i contains two disjoint types of information for the

---

[*]Also affiliated to Collegio Carlo Alberto, Torino, Italy

MSC 2010 subject classifications: Primary 62G05 ; Secondary 62C20

Keywords and phrases: disclosure risk assessment, microdata sample, nonparametric inference, optimal minimax procedure, Poisson abundance model, polynomial approximation

i-th individual: identifying information and sensitive information. Identifying information consists of the values of a set of categorical variables, which might be matchable to known units of the population. A risk of disclosure arises from the possibility that an intruder might succeed in identifying a microdata unit through such a matching and hence be able to disclose the sensitive information on this unit. To quantify the risk of disclosure, microdata sample records are cross-classified according to potentially identifying variables, i.e., $\boldsymbol{X}(n)$ is partitioned in $K_n \leq n$ cells with corresponding frequency counts $(Y_1(\boldsymbol{X}, n), \ldots, Y_{K_n}(\boldsymbol{X}, n))$ such that $\sum_{1\leq i\leq K_n} Y_j(\boldsymbol{X}, n) = n$, where $Y_j(\boldsymbol{X}, n)$ denotes the frequency of the j-th cell out of the sample $\boldsymbol{X}(n)$. A risk of disclosure arises from cells in which both sample frequencies and population frequencies are small. Of special interest are cells with frequency 1 (singletons or uniques) since, assuming no errors in the matching process or data sources, for these cells the match is guaranteed to be correct. This has motivated inference on measures of disclosure risk that are functionals of the number of singletons, the most common being the number τ_1 of sample singletons which are also population singletons. See, e.g., Bethlehem et al. [2] and Skinner et al. [23] for a thorough discussion on measures of disclosure risk.

The Poisson abundance model is arguably the most natural, and weak, modeling assumption to infer τ_1. If $\bar{n} = n + \lambda n$, with $\lambda > 0$, it assumes that: i) the population records $(X_1, \ldots, X_{n+\lambda n})$ can be ideally extended to a sequence $\boldsymbol{X} := (X_i)_{i\geq 1}$, of which $\boldsymbol{X}(n)$ is an observable subsample; ii) the X_i's are independent and identically distributed according to an unknown distribution $(p_j)_{j\geq 1}$, where p_j is the probability of the j-th cell in which $\boldsymbol{X}$ may be cross-classified; iii) the sample size is a Poisson random variable N with mean n, in symbols $N \sim \mathrm{Poiss}(n)$. Then sample records $\boldsymbol{X}(N) = (X_1, \ldots, X_N)$ result in K_N cells with frequencies $(Y_1(\boldsymbol{X}, N), \ldots, Y_{K_N}(\boldsymbol{X}, N))$ such that $Y_j(\boldsymbol{X}, N) \sim \mathrm{Poiss}(np_j)$ for $j = 1, \ldots, K_N$, $Y_{j_1}(\boldsymbol{X}, N)$ is independent of $Y_{j_2}(\boldsymbol{X}, N)$ for any $j_1 \neq j_2$, and $\sum_{1\leq j\leq K_N} Y_j(\boldsymbol{X}, N) = N$. As discussed in Section 2.4 of Skinner and Elliot [22], nonparametric estimation of τ_1 under the Poisson abundance model is an intrinsically difficult problem. It shares the well-known difficulties of the classical problem of estimating the number of unseen species (e.g., Good and Toulmin [10], Efron and Thisted [8], Orlitsky et al. [17]). In particular, nonparametric estimators of τ_1 may be "very unreasonable" since they are subject to serious upward bias and high variance for small sampling fractions of the population, i.e. for $(\lambda + 1)^{-1} < 1/2$. To overcome these issues, in the last three decades stronger modeling assumptions have been considered. These studies resulted in a range of parametric and semiparametric approaches, both frequentist and Bayesian, to infer τ_1, e.g., Bethlehem et al. [2], Samuels [21], Skinner and Elliot [22], Reiter [18], Rinott and Shlomo [19], Skinner and Shlomo [24], Manrique-Vallier and Reiter [13], Manrique-Vallier and Reiter [14], Carota et al. [4] and Carota et al. [5].

In this paper, we first study nonparametric estimation of τ_1 under the Poisson abundance model for sample records. Given a collection of sample records $(X_1, \ldots, X_n)$ from the population $(X_1, \ldots, X_{n+\lambda n})$, we introduce a class of nonparametric linear estimators of τ_1 that are simple, computationally efficient and scalable to massive datasets. We show that our estimators admit an interpretation as (smoothed) nonparametric empirical Bayes estimators in the sense of Robbins [20], and we prove theoretical guarantees for them that hold uniformly for any distribution $(p_j)_{j\geq 1}$. In particular, we show that the proposed estimators provably estimate τ_1 all of the way up to the sampling fraction $(\lambda+1)^{-1} \propto (\log n)^{-1}$, with vanishing normalized mean-square error (NMSE) as n becomes large. Then, by relying on recent techniques developed in Wu and Yang [29] in the context of optimal estimation of the support size of discrete distributions, we establish a lower bound for the minimax NMSE for the estimation of τ_1. This result allows us to show that $(\lambda+1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population, and that estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for large n. This is the main result of our paper, and it provides a precise answer to the question raised by Skinner and Elliot [22] about the feasibility of nonparametric estimation of τ_1 under the Poisson abundance model and for a sampling fraction $(\lambda+1)^{-1} < 1/2$. Indeed our result shows that nonparametric estimation of τ_1 has uniformly provable guarantees, in terms of vanishing NMSE for large n, if and only if $(\lambda+1)^{-1} \propto (\log n)^{-1}$.

The paper is structured as follows. In Section 2 we introduce a class of nonparametric estimators for τ_1, and we show that they provably estimate τ_1 all of the way up to the sampling fraction $(\lambda+1)^{-1} \propto (\log n)^{-1}$, with vanishing NMSE as n becomes large. In Section 3 we show that $(\lambda+1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population, and that estimators' NMSE is near optimal for large n. Section 4 contains a numerical illustration of the proposed estimators. Proofs and deferred to the Appendix.

2. A nonparametric estimator of $\boldsymbol{\tau_1}$. We consider an infinite sequence of observations $\boldsymbol{X}$, and we assume that $\boldsymbol{X}(N) = (X_1, \ldots, X_N)$ is the microdata sample of random size N under the Poisson abundance model. We suppose that $\boldsymbol{X}(N)$ is a subsample of $(X_1, \ldots, X_{M+N})$, where

$M \sim \mathrm{Poiss}(\lambda n)$, with $\lambda > 0$ and independent of N. In the present framework $(X_{N+1}, \ldots, X_{N+M})$ may be seen as the unobservable population. When the sample records are cross-classified according to the potentially identifying variables, the sample $(X_1, \ldots, X_N)$ is partitioned in $K_N \leq N$ cells with corresponding frequency counts $(Y_1(\boldsymbol{X}, N), \ldots, Y_{K_N}(\boldsymbol{X}, N))$ such that $\sum_{1 \leq i \leq K_N} Y_j(\boldsymbol{X}, N) = N$. Hereafter we denote by $Z_i(\boldsymbol{X}, N)$ the number of cells with frequency i, and by $Z_{\bar{i}}(\boldsymbol{X}, N)$ the number of cells with frequency greater or equal than i, for any index $i \geq 1$. We are interest in estimating the number τ_1 of sample uniques which are also population uniques, namely the functional

$$\tau_1(\mathbf{X}, N, M) = \sum_{j \geq 1} \mathbb{1}_{\{Y_j(\boldsymbol{X}, N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X}, N+M)=1\}}.$$

We recall that the frequency counts $Y_j(\boldsymbol{X}, N)$'s are independent, and that they are Poisson distributed with parameter np_j, where p_j is the unknown probability associated to the j-th cell, that is $p_j \in [0, 1]$ for $j \geq 1$ such that $\sum_{j \geq 1} p_j = 1$. We will denote by $\boldsymbol{Y}(\boldsymbol{X}, N) := (Y_1(\boldsymbol{X}, N), \ldots)$ the whole sequence of the cell's frequency count, when we are provided with a sample of size N.

To fix the notation, in the sequel we will write $f \lesssim g$, for two generic functions f and g, iff there exists a universal constant $C > 0$ such that $f(x) \leq Cg(x)$; we will further write $f \asymp g$ whenever both $f \lesssim g$ and $g \lesssim f$ are satisfied. Let us denote by $\mathscr{P}$ the set of all possible distributions over $\mathbb{N}$, i.e. $\mathscr{P} := \{P = \sum_{j \geq 1} p_j \delta_j : \ p_j \in [0, 1], \text{ with} \sum_{j \geq 1} p_j = 1\}$, where δ_j denotes the Dirac measure centered at $j \in \mathbb{N}$. An estimator of $\tau_1(\boldsymbol{X}, N, M)$ is understood to be a measurable function $\hat{\rho}_1(\boldsymbol{X}(N), N)$ depending on the available sample $\boldsymbol{X}(N)$ and the actual size of the observed sample N. We will evaluate the performance of a generic estimator $\hat{\rho}_1(\boldsymbol{X}(N), N)$ of $\tau_1(\boldsymbol{X}, N, M)$, by its worst–case NMSE, defined as

$$\mathscr{E}_{\lambda,n}(\hat{\rho}_1(\boldsymbol{X}(N), N)) := \sup_{P \in \mathscr{P}} \frac{\mathbb{E}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2]}{n^2}, \tag{1}$$

where $\mathbb{E}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2]$ is the mean squared error (MSE) of $\hat{\rho}_1$, also denoted by $\mathrm{MSE}[\hat{\rho}_1(\boldsymbol{X}(N), N)]$. The use of the NMSE (1) has been recently proposed in Orlitsky et al. [17] in the context of the estimation of the number of unseen species.

A nonparametric estimator for $\tau_1(\boldsymbol{X}, N, M)$ may be deduced comparing expectations, indeed it is easy to see that:

$$\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] = \sum_{i \geq 0} (-1)^i \lambda^i (i+1) \mathbb{E}[Z_{i+1}(\boldsymbol{X}, N)] \tag{2}$$

from which we may define the following estimator

$$\hat{\tau}_1(\boldsymbol{X}(N), N) = \sum_{i\geq 0}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X}, N), \tag{3}$$

which turns out to be unbiased by construction. See Appendix A.1 for the determination of (2). The estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ admits a natural interpretation as a nonparametric empirical Bayes estimator in the sense of Robbins [20]. More precisely, $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is the posterior expectation of $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$ with respect to an unknown prior distribution on the p_i's that is estimated from the $Y_j(\mathbf{X}, N)$. See Appendix A.2 for details. The next theorem legitimates the use of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ as an estimator of $\tau_1(\boldsymbol{X}, N, M)$, for $\lambda < 1$, i.e. when the size of the unobserved population is less or equal than n, the size of the observed sample.

THEOREM 1. *For any positive real numbers x and y let $\lfloor x\rfloor$ denote the integer part of x and let $x \vee y$ denote the maximum between x and y. If $\lambda < 1$, for any $P \in \mathscr{P}$, we get*

$$\mathbb{E}[\hat{\tau}_1(\boldsymbol{X}(N), N)] = \mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] = \sum_{j\geq 1} np_j e^{-(\lambda+1)np_j} \tag{4}$$

and

$$\begin{aligned} &\mathrm{Var}[\tau_1(\boldsymbol{X}, N, M) - \hat{\tau}_1(\boldsymbol{X}(N), N)] \\ &\qquad \leq \Psi^2(\lambda)\mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1}, \end{aligned} \tag{5}$$

where in (5) we defined $\Psi(\lambda) = (j^*+1)\lambda^{j^*}$ such that $j^* = \lfloor(2\lambda-1)/(1-\lambda)\rfloor \vee 0$.

See Appendix A.3 for the proof of Theorem 1. According to Theorem 1, for $\lambda < 1$ one has $\mathbb{E}[\hat{\tau}_1(\boldsymbol{X}(N), N)] = \mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$ and $\mathrm{Var}[\tau_1(\boldsymbol{X}, N, M) - \hat{\tau}_1(\boldsymbol{X}(N), N)] \lesssim n$ upon noticing that $\mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] \leq \mathbb{E}[N] = n$. That is, in expectation, $\hat{\tau}_1(\boldsymbol{X}(N), N)$ approximate $\tau_1(\boldsymbol{X}, N, M)$ to within n. Hence we formalize our considerations in the following.

COROLLARY 1. *Assume that $\lambda < 1$, then the nonparametric estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ defined in (3) satisfies*

$$\mathscr{E}_{\lambda,n}(\hat{\tau}_1(\boldsymbol{X}(N), N)) \lesssim \frac{1}{n} \tag{6}$$

for any $n \geq 1$.

This legitimates the use of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ as an estimator of $\tau_1(\boldsymbol{X}, N, M)$ under the hypothesis $\lambda < 1$, which unfortunately is a quite restrictive assumption within the framework of disclosure risk: indeed the size of the unobserved sample is usually much bigger than the size of the available one. However the derivation of a variance bound for $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is a crucial step for our study. Indeed it reveals that the assumption $\lambda < 1$ is necessary to obtain a finite estimate of the variance. This variance issue of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is determined by the geometrically increasing magnitude of the coefficients $(i+1)(-\lambda)^i$. Indeed, as $\lambda \geq 1$, the estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ grows superlinearly as $(i+1)(-\lambda)^i$ for the largest i such that $Z_{i+1}(\boldsymbol{X}, N) > 0$, thus eventually far exceeding $\tau_1(\boldsymbol{X}, N, M)$ that grows at most linearly. This is the main reason why $\hat{\tau}_1(\boldsymbol{X}(N), N)$ become useless for $\lambda \geq 1$, thus requiring an adjustment via suitable smoothing techniques. Hereafter we follow ideas originally developed by Good and Toulmin [10], Efron and Thisted [8] and Orlitsky et al. [17] for nonparametric estimators of the number of unseen species. Specifically, we propose a smoothed version of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ by truncating the series (3) at an independent random location L and averaging over the distribution of L, i.e.,

$$
\begin{aligned}
\hat{\tau}_1^L(\boldsymbol{X}(N), N) &= \mathbb{E}_L\left[\sum_{i=1}^{L}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X}, N)\right] \\
&= \sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L \geq i) Z_{i+1}(\boldsymbol{X}, N).
\end{aligned} \tag{7}
$$

For any $\lambda \geq 1$, as the the index i in (7) increases, the tail probability $\mathbb{P}[L \geq j]$ compensate for the exponential growth of $(i+1)(-\lambda)^i$, thereby stabilizing the variance. In the next theorem we show that for $\lambda \geq 1$ the estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is biased for $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$, and we provide a bound for the MSE of $\hat{\tau}_1(\boldsymbol{X}(N), N)$.

THEOREM 2. *Suppose that $\lambda \geq 1$, then $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is a biased estimator of $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$ with*

$$
\begin{aligned}
&\mathbb{E}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \\
&\quad = \mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] + \sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s.
\end{aligned} \tag{8}
$$

and

$$
\begin{aligned}
&\mathrm{MSE}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \\
&\quad \le \left(\sum_{j\ge 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s \right)^2 \\
&\qquad + (\mathbb{E}_L[(L+1)\lambda^L])^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1}.
\end{aligned} \tag{9}
$$

Choosing different smoothing distributions for the random variable L yields different estimators for $\tau_1(\boldsymbol{X}, N, M)$. Following Orlitsky et al. [17], three possible choices for the distribution of L are the following: i) a Poisson distribution with parameter $\beta > 0$; ii) a Binomial distribution with parameter $(x_0, (\lambda+1)^{-1})$; iii) a Binomial distribution with parameter $(x_0, 2/(\lambda+2))$. In particular, it can be shown that the choice of the Binomial distribution with parameter $(x_0, (\lambda+1)^{-1})$ corresponds to the truncation at the point x_0 of the Euler transformation of the estimator (3). To choose the parameter β of the Poisson distribution and the parameter x_0 of the Binomial distribution, one should look for $\tilde{\beta}$ and $\tilde{x}_0$ which minimizes the MSE bound (9). Once the values of $\tilde{\beta}$ and $\tilde{x}_0$ are determined explicitly, we are able to obtain limit of predictability for $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$. That is, for some $\delta > 0$ we are able to specify the maximum value of the sampling fraction λ for which $\mathscr{E}_{\lambda,n}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) < \delta$. This gives a provable (performance) guarantee for the estimation of $\tau_1(\boldsymbol{X}, N, M)$ in terms of the sampling fraction λ. The next proposition specifies the limit of predictability for the estimator under the choice of a Poisson distribution with parameter β for the smoothing distribution L.

PROPOSITION 1. *Let L be a Poisson random variable with parameter β. Then*

$$
\mathrm{MSE}[\hat{\tau}_1^L(\boldsymbol{X}(N), N)] \le e^{-2\beta} n^2 + n e^{2\beta(2\lambda-1)} \tag{10}
$$

whose upper bound is minimized when

$$
\tilde{\beta} = \frac{1}{4\lambda} \log\left(\frac{n}{2\lambda - 1}\right).
$$

for any $\lambda \ge 1$. Moreover, if L is a Poisson random variable with parameter $\tilde{\beta}$ then

$$
\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) \le \frac{A(\lambda)}{n^{1/(2\lambda)}}, \tag{11}
$$

and for any $\delta \in (0,1)$

$$\lim_{n\to+\infty} \frac{\max\left\{\lambda :\ \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N),N)) \leq \delta\right\}}{\log(n)} \geq \frac{1}{2\log(A/\delta)} \tag{12}$$

where $A(\lambda)$ *is continuous in* $[1,+\infty)$ *with* $\lim_{\lambda\to+\infty} A(\lambda) = 1$ *and* $A = \max_{\lambda\geq 1} A(\lambda) < +\infty$.

See Appendix A.5 for the proof of Proposition 1. Similar results hold true when L is assumed to be a Binomial random variable: the derivation of these results follows along similar lines as the proof of Proposition 1. Hence we state the following result in presence of a Binomial smoothing without proof.

PROPOSITION 2. *Let* L *be a Binomial random variable with parameter* $(x_0, 2/(\lambda+2))$. *Then*

$$\begin{aligned} &\mathrm{MSE}[\hat{\tau}_1^L(\boldsymbol{X}(N),N)] \\ &\qquad \leq n\left(\frac{\lambda}{\lambda+2}\right)^{2x_0}\left[3^{10x_0/3} + n\left(\frac{\lambda}{2(\lambda+1)}\right)^2\right] \end{aligned} \tag{13}$$

whose upper bound is minimized when

$$\tilde{x}_0 = \left\lfloor \frac{3}{10}\log_3\left(n\frac{\lambda^2}{(\lambda+1)(\lambda^2(3^{10/3}-1)-4\lambda-4)}\right)\right\rfloor$$

for any $\lambda \geq 1$. *Moreover, if* L *is a Binomial random variable with parameter* $(\tilde{x}_0, 2/(\lambda+2))$ *then*

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N),N)) \leq \frac{C(\lambda)}{n^{3\log_3(1+2/\lambda)/5}}, \tag{14}$$

and for any $\delta \in (0,1)$

$$\lim_{n\to+\infty} \frac{\max\left\{\lambda :\ \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta\right\}}{\log(n)} \geq \frac{6}{5\log(3)\log(C/\delta)} \tag{15}$$

where $C(\lambda)$ *is continuous in* $[1,+\infty)$ *with* $\lim_{\lambda\to+\infty} C(\lambda) = 1$ *and* $C = \max_{\lambda\geq 1} C(\lambda)$.

3. Optimality of the proposed estimators. In Section 2 we have defined two different estimators of $\tau_1(\boldsymbol{X}, N, M)$ providing guarantees of their performance as $n \to +\infty$ in terms of the NMSE. We have already remarked that the case $\lambda \geq 1$ is the most interesting one for estimating the disclosure risk $\tau_1(\boldsymbol{X}, N, M)$, indeed the fraction of the unobserved sample λ is usually much larger than 1. Throughout the section we assume that $\lambda \geq 1$ and we prove that the proposed estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is essentially optimal. More precisely we determine a lower bound for the best worst–case NMSE, defined by

$$\mathscr{E}(\lambda, n) := \inf_{\hat{\rho}_1} \mathscr{E}_{\lambda,n}(\hat{\rho}_1(\boldsymbol{X}(N), N)) \tag{16}$$

where the infimum in the previous definition runs over all possible estimators of $\tau_1(\boldsymbol{X}, N, M)$. We will then see that the determined lower bound essentially matches with the upper bound (11). In the sequel we refer to $\mathscr{E}(\lambda, n)$ as the *(normalized) minimax risk*.

The theorem we are going to state below provides us with a lower bound for the minimax risk.

THEOREM 3. *Assume that $1 + \lambda > e^2$. Then, there exists a universal constant $K > 0$ such that for any n sufficiently big we have*

$$\mathscr{E}(\lambda, n) \geq K \cdot \begin{cases} 1 & \text{if } \lambda + 1 > \log(n) \\ \frac{1+\lambda}{\log(n)} \left(\frac{\sqrt{\log(n)}}{n(1+\lambda)} \right)^{e^2/(1+\lambda)} & \text{if } \lambda + 1 \leq \log(n) \end{cases} \tag{17}$$

From Theorem 3, it is clear that the minimax risk goes to zero if $\lambda + 1 = o(\log(n))$ and the rate is provided by the following Corollary.

COROLLARY 2. *Assume that $1 + \lambda > e^2$, then there exist universal constants $c > 0$ and $c' > 0$ such that for any n sufficiently large*

$$\mathscr{E}(\lambda, n) \geq c \frac{1}{n^{c'/\lambda}}. \tag{18}$$

Corollary 2 is an easy consequence of Theorem 3, indeed, when $\lambda + 1 > \log(n)$ the two lower bounds in (17)–(18) are constants, whereas if $\lambda + 1 \leq \log(n)$ it is easy to observe that the leading term in (17) (as $n \to +\infty$) is of order $1/n^{c'/\lambda}$ as in (18) for some $c' > 0$. Corollary 2 provides us with a lower bound for the NMSE of any estimator of the disclosure risk $\tau_1(\boldsymbol{X}, N, M)$. The lower bound (18) has an important implication: without imposing any parametric assumption on the model, one can estimate $\tau_1(\boldsymbol{X}, N, M)$ with vanishing NMSE all the way up to $\lambda \propto \log n$. It is then impossible to determine an estimator having provable guarantees (in terms of vanishing NMSE) when $\lambda = \lambda(n)$ goes to $+\infty$ much faster than $\log(n)$, as a function of n. By the limit of predictability (12) determined for the estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$, we conclude that the proposed estimator is optimal, because its limit of predictability matches (asymptotically) with its maximum possible value $\lambda \propto \log(n)$.

3.1. *Guideline for the proof of Theorem 3.* In the present section we provide the main ingredients for the proof of Theorem 3, technical results and related proofs are deferred to the Appendix. In the sequel we will write $\mathbb{E}_P^n$ to make explicit the dependence of the expected value w.r.t. P and the parameter n of the Poisson random variable N.

The starting point for the proof of Theorem 3 is the next Lemma 1, which is an interesting result in its own right and will help a lot in the proof of Theorem 3. Remark that the definition of the minimax risk in (16) allows for estimators depending on the whole sample $\boldsymbol{X}(N)$, while $\tau_1(\boldsymbol{X}, N, M)$ depends only on the frequencies $\boldsymbol{Y}(\boldsymbol{X}, N+M)$ and $\boldsymbol{Y}(\boldsymbol{X}, N)$. Thus, we feel like there should be no gain of information in using estimators depending on $\boldsymbol{X}(N)$ over estimators depending only on the frequencies $\boldsymbol{Y}(\boldsymbol{X}, N)$. This is made formal in the next lemma, proved in Section B.1. Note that this is convenient since $(\boldsymbol{X}, k) \mapsto \boldsymbol{Y}(\boldsymbol{X}, k)$ is nicely distributed under the Poisson model.

LEMMA 1. *The following equality is true*

$$\mathscr{E}(\lambda, n) = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N))^2],$$

where the infinimum in the previous equation is understood to be taken with respect to all measurable maps $\hat{\rho} : \mathbb{N}^{\mathbb{N}} \to \mathbb{R}$.

The next step is to use Jensen's inequality to deduce that

$$\begin{aligned}\mathscr{E}(\lambda, n) &= \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[\mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2 \mid \boldsymbol{Y}(\boldsymbol{X}, N)]] \\ &\geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\mathbb{E}_P^n[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2]\end{aligned}$$

Note that there is no explicit dependency on $\boldsymbol{X}$ and M anymore in the last display, but only on the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$ which, under P, is distributed as an infinite vector of independent Poisson random variables with parameters $(np_1, np_2, \dots)$. Besides observe also that $N = \sum_{j \geq 1} Y_j(\boldsymbol{X}, N)$. For the sake of notational simplicity, in the sequel $\boldsymbol{Y}$ will stand for the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$, and we also let

$$\tilde{\tau}_1(\boldsymbol{Y}, P, n) := \mathbb{E}_P^n[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N)]$$

$$= \sum_{j\geq 1} \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{E}_P^n[\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)-Y_j(\boldsymbol{X},N)=0\}} \mid \boldsymbol{Y}(\boldsymbol{X},N)].$$

Remark that $(Y_j(\boldsymbol{X},N+M)-Y_j(\boldsymbol{X},N) : j \in \mathbb{N})$ is independent of $\boldsymbol{Y}(\boldsymbol{X},N)$ and is a collection of independent Poisson random variables with intensities $(\lambda n p_j : j \in \mathbb{N})$. Henceforth, we get

$$\tilde{\tau}_1(\boldsymbol{Y},P,n) = \sum_{j\geq 1} e^{-\lambda n p_j} \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}, \tag{19}$$

and besides

$$\mathscr{E}(\lambda,n) \geq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^n[(\tilde{\tau}_1(\boldsymbol{Y},P,n)-\hat{\rho}(\boldsymbol{Y}))^2]. \tag{20}$$

We now trade $\tilde{\tau}_1(\boldsymbol{Y},P,n)$ for its expectation. Let us introduce $\bar{\tau}_1(P,n) := \mathbb{E}_P^n[\tilde{\tau}_1(\boldsymbol{Y},P,n)]$. Recall that under P the vector $\boldsymbol{Y}$ is distributed as independent Poisson with parameters $(np_1, np_2, \dots)$. Hence,

$$\bar{\tau}_1(P,n) = \sum_{j\geq 1} e^{-\lambda n p_j} \mathbb{E}_P^n[\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}] = n\sum_{j\geq 1} p_j e^{-(1+\lambda)np_j}.$$

Similarly, for any $P \in \mathscr{P}$,

$$\mathrm{Var}(\tilde{\tau}_1(\boldsymbol{Y},P,n)) = \sum_{j\geq 1} np_j e^{-(1+2\lambda)np_j}\big\{1-np_je^{-np_j}\big\} \leq n. \tag{21}$$

Thus from (20), Young's inequality, we find that

$$\mathscr{E}(\lambda,n) \geq \frac{1}{2}\inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} n^{-2}\mathbb{E}_P^n[(\bar{\tau}_1(P,n)-\hat{\rho}(\boldsymbol{Y}))^2] - n^{-1}. \tag{22}$$

The remainder of the proof mostly follows the reduction scheme used in Wu and Yang [28, 29] which consists on reducing the problem to finding the best polynomial approximation (in uniform norm) to a suitable function.

The first step of the reduction scheme is to trade $\mathscr{P}$ in (16) for a slightly more convenient set. We let $S \in \mathbb{N}$ be the only integer satisfying

$$n(1+\lambda) \leq S \leq n(1+\lambda)+1, \tag{23}$$

and we also let, for some constant $c_0 > 0$ to be determined later,

$$\xi := (2c_0/e)\min\{(1+\lambda)\log n,\, \log^2 n\}. \tag{24}$$

Then, for another constant $c_1 > 0$ and for $\varepsilon > 0$ to be determined later, we define

$$\mathscr{P}' := \Big\{ \textstyle\sum_{k=1}^S p_k \delta_k : \ p_k \in [0, \xi S^{-1}], \ |\sum_{k=1}^S p_k - 1| \leq c_1 \varepsilon/\xi \Big\}. \tag{25}$$

Remark that $\mathscr{P}'$ contains measures that are not probability measures, and hence it is not clear a priori that we can lower bound the supremum over $\mathscr{P}$ by the supremum over $\mathscr{P}'$. The next proposition shows that it is fine as long as $c_1\varepsilon$ is not too large. Here and after, under $P \in \mathscr{P}'$, $\boldsymbol{Y}$ is understood as a vector of independent Poisson random variables with intensities $(np_1, \dots, np_S, 0, \dots)$, with $\sum_{j=1}^S p_j$ not necessarily equal to one, and $P \mapsto \bar{\tau}_1(P)$ is extended trivially from $\mathscr{P}$ to $\mathscr{P}'$ by letting $\bar{\tau}_1(P, n) := n \sum_{j=1}^S p_j e^{-n(1+\lambda)p_j}$, $P \in \mathscr{P}'$. The next proposition is proved in Section B.2.

PROPOSITION 3. *Assume that $c_1\varepsilon = o(\xi)$ as $n \to \infty$ and let define $n' := n(1 + c_1\varepsilon/\xi)$. Then as $n \to \infty$,*

$$\begin{aligned}\mathscr{E}(\lambda, n) &\geq \frac{1}{4} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} n^{-2} \mathbb{E}_P^{n'}[(\bar{\tau}_1(P, n) - \hat{\rho}(\boldsymbol{Y}))^2] - (n^{-1} + 9c_1^2\varepsilon^2/2) \\ &\geq \frac{\varepsilon^2}{4} \Big\{ \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{P}_P^{n'}\big(|\bar{\tau}_1(P, n) - \hat{\rho}(\boldsymbol{Y})| > n\varepsilon\big) - 18c_1^2 \Big\} - n^{-1}.\end{aligned}$$

We are now in position to lower bound the risk by the Bayes risk. To do so, we follow the prior construction of Wu and Yang [28, 29]. For some $L \in \mathbb{N}$ to be determined later, but satisfying $L \leq c_2\xi$ for some constant $c_2 > 0$, we let U and V be two random variables taking values in $[0, \xi S^{-1}]$ such that when n is large enough,

$$\begin{gathered}\mathbb{E}[U^k] = \mathbb{E}[V^k] \quad \forall k \in \{0, \dots, L+1\}, \\ \mathbb{E}[U] = \mathbb{E}[V] = S^{-1}, \quad \mathrm{Var}(U) \leq \xi S^{-2}, \quad \mathrm{Var}(V) \leq \xi S^{-2}, \\ \mathbb{E}[U e^{-n(1+\lambda)U}] \geq \mathbb{E}[V e^{-n(1+\lambda)V}] + S^{-1} K \min\big\{1, \sqrt{\xi/L^2}\exp(-L^2/\xi)\big\}\end{gathered}$$

The existence of such random variables is guaranteed by Theorem C.1 for a universal constant $K > 0$. Then we let $\boldsymbol{U} := (U_1, \dots, U_S)$, respectively $\boldsymbol{V} := (V_1, \dots, V_S)$, be an independent vector of i.i.d. copies of U, respectively V. Denoted by $\mathcal{M}(\mathbb{N})$ the space of all measures on $\mathbb{N}$, we construct the following random variable $Q : [0,1]^S \to \mathcal{M}(\mathbb{N})$ such that $Q(\boldsymbol{U}) := \sum_{k=1}^S U_k \delta_k$. Then, from the Proposition 3 and Hölder's inequality, we find that $\mathscr{E}(\lambda, n)$ is bounded from below by $-n^{-1}$ plus

$$\frac{\varepsilon^2}{4}\Big\{ \inf_{\hat\rho} \Big(\frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{U})}\big(|\bar\tau_1(Q(\boldsymbol{U}),n) - \hat\rho(\boldsymbol{Y})| > n\varepsilon\big)\mathbb{1}_{\mathscr{P}'}(Q(\boldsymbol{U}))\Big] \\ + \frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{V})}\big(|\bar\tau_1(Q(\boldsymbol{V}),n) - \hat\rho(\boldsymbol{Y})| > n\varepsilon\big)\mathbb{1}_{\mathscr{P}'}(Q(\boldsymbol{V}))\Big]\Big) - 18c_1^2\Big\},$$

which is in turn lower bounded by

$$\frac{\varepsilon^2}{4}\Big\{ \inf_{\hat\rho} \Big(\frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{U})}\big(|\bar\tau_1(Q(\boldsymbol{U}),n) - \hat\rho(\boldsymbol{Y})| > n\varepsilon\big)\Big] \\ + \frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{V})}\big(|\bar\tau_1(Q(\boldsymbol{V}),n) - \hat\rho(\boldsymbol{Y})| > n\varepsilon\big)\Big]\Big) \\ - 18c_1^2 - \frac{1}{2}\mathbb{P}(Q(\boldsymbol{U}) \notin \mathscr{P}') - \frac{1}{2}\mathbb{P}(Q(\boldsymbol{V}) \notin \mathscr{P}')\Big\} - n^{-1}.$$

The last display follows because we don't have $Q(\boldsymbol{U})$ nor $Q(\boldsymbol{V})$ almost-surely in $\mathscr{P}'$, but it is clear that the strong law of large numbers implies they should be concentrated on $\mathscr{P}'$. Formally, an application of Bernstein's inequality (see Section B.3 below) leads to the following proposition.

PROPOSITION 4. *Assume that $c_1\varepsilon = o(\xi)$ as $n \to \infty$. Then, there exists a constant $C > 0$, depending only on c_1, such that for n large enough,*

$$\varepsilon^2 \geq \frac{C\xi^3}{n(1+\lambda)} \Longrightarrow \mathbb{P}\big(Q(\boldsymbol{U}) \notin \mathscr{P}'\big) \leq c_1^2.$$

Thus under the conditions of Proposition 4, we get

$$\mathscr{E}(\lambda,n) \geq \frac{\varepsilon^2}{4}\Big\{ \inf_{\hat\rho} \Big(\frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{U})}\big(|\bar\tau_1(Q(\boldsymbol{U}),n) - \hat\rho(\boldsymbol{Y})| > n\varepsilon\big)\Big] \\ + \frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{V})}\big(|\bar\tau_1(Q(\boldsymbol{V}),n) - \hat\rho(\boldsymbol{Y})| > n\varepsilon\big)\Big]\Big) - 19c_1^2\Big\} - n^{-1}. \tag{26}$$

We now wish to trade $\bar\tau_1(Q(\boldsymbol{U}),n)$ and $\bar\tau_1(Q(\boldsymbol{V}),n)$ for their expectations in the last equation. Intuitively, this should not be problematic since they are sums of i.i.d. random variables, they should concentrate near their expectations for S large enough. We made this formal using a Hoeffding argument in the next proposition, proved in Section B.4.

PROPOSITION 5. *Let everything as above. Then,*

$$\varepsilon^2 \geq \frac{2\xi\log(1/c_1^2)}{n(1+\lambda)} \Longrightarrow \mathbb{P}\big(|\bar\tau_1(Q(\boldsymbol{U}),n) - \mathbb{E}[\bar\tau_1(Q(\boldsymbol{U}),n)]| > n\varepsilon/2\big) \leq 2c_1^2.$$

Obviously, Proposition 5 is also true for $\bar{\tau}_1(Q(\boldsymbol{V}), n)$. We now assume that the conditions of Propositions 3, 4 and 5 are met, and thus we obtain from (26) that

$$\begin{aligned}\mathscr{E}(\lambda, n) \geq \frac{\varepsilon^2}{4}\Big\{ \inf_{\hat{\rho}} \Big(\frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{U})}\big(|\mathbb{E}[\bar{\tau}_1(Q(\boldsymbol{U}), n)] - \hat{\rho}(\boldsymbol{Y})| > n\varepsilon/2\big)\Big] \\ + \frac{1}{2}\mathbb{E}\Big[\mathbb{P}^{n'}_{Q(\boldsymbol{V})}\big(|\mathbb{E}[\bar{\tau}_1(Q(\boldsymbol{V}), n)] - \hat{\rho}(\boldsymbol{Y})| > n\varepsilon/2\big)\Big]\Big) - 21c_1^2\Big\} - n^{-1}.\end{aligned}$$

Now remark that,

$$\mathbb{E}[\bar{\tau}_1(Q(\boldsymbol{U}), n)] = nS\mathbb{E}[U \exp\{-n(1+\lambda)U\}],$$

and besides observe that whenever n is large enough, we will have $C\xi^2 \geq 2\log(1/c_1^2)$. We furthermore assume that ε^2 satisfies

$$\max\Big\{\frac{8}{nc_1^2}, \frac{C\xi^3}{n(1+\lambda)}\Big\} \leq \varepsilon^2 \leq K^2 \min\big\{1,\ (\xi/L^2)\exp(-2L^2/\xi)\big\}. \tag{27}$$

It is not clear yet that (27) can be satisfied, we claim this is the case and we delay the proof of the claim at the end of the section. When the claim is true, we pick ε to be equal to the r.h.s. of (27). Then, the previous computations and the classical Le Cam method with two hypothesis imply that

$$\begin{aligned}&\mathscr{E}(\lambda, n) \\ &\quad \geq \frac{\varepsilon^2}{8}\Big\{1 - \mathsf{TV}\Big(\otimes_{j=1}^S \mathbb{E}[\mathrm{Poiss}(n'U_j)], \otimes_{j=1}^S \mathbb{E}[\mathrm{Poiss}(n'V_j)]\Big) - 42c_1^2\Big\} - \frac{1}{n} \\ &\quad \geq \frac{\varepsilon^2}{8}\Big\{1 - S\cdot\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) - 43c_1^2\Big\}.\end{aligned}$$

We now make explicit our choice for the value of L. We need to choose it small enough such that $\mathbb{E}[U\exp\{-n(1+\lambda)U\}]$ and $\mathbb{E}[V\exp\{-n(1+\lambda)V\}]$ are maximally separated, but also large enough such that the previous display is non-negative. For a constant $c_3 > 0$ to be chosen accordingly later, and for $A(\lambda, n) > 0$ solution of

$$A(\lambda, n)\log A(\lambda, n) = c_0^{-1} + c_0^{-1}\frac{\log(1+\lambda) - (1/2)\log\log(n) + \log(c_3)}{\log(n)},$$

we pick L to be the smallest integer satisfying the bound,

$$L \geq \begin{cases} 2c_0\log(n) & \text{if } 1+\lambda > \log(n), \\ c_0 A(\lambda, n)\log(n) & \text{if } 1+\lambda \leq \log(n). \end{cases} \tag{28}$$

Remark that this choice ensure that $L \leq c_2\xi$ for some constant $c_2 > 0$, as requested. Without loss of generality we further assume that $L + 2 \leq c_2\xi$. We are then able to state the following propositon, which will be proved in Section B.5.

PROPOSITION 6. *Assume that $c_1 c_2 \varepsilon \leq 1$. Then, the constant $c_3 > 0$ can be chosen depending only on the choice of c_0 and such that for n large enough,*

$$S\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \leq \frac{1}{2}.$$

Now observe that $\varepsilon \leq K$ by (27), thus choosing $c_1 = \min\{1/\sqrt{172},\ 1/(c_2 K)\}$, we get $c_1 c_2 \varepsilon \leq 1$ and $43 c_1^2 \leq 1/4$. Together with the last proposition, this implies that $\mathscr{E}(\lambda, n) \geq \varepsilon^2/32$, at least if ε can be chosen to satisfy Equation (27).

We now prove that (27) is satisfied. When $1+\lambda > \log(n)$, then the r.h.s. of (27) is always greater than a constant, while the the l.h.s. is always smaller than $O(1/n)$. Thus in this case we take ε^2 to be a suitable constant, and (27) is satisfied, giving $\mathscr{E}(\lambda, n) \gtrsim 1$.

When $1 + \lambda \leq \log(n)$, then there is always a constant $K' > 0$ such that the r.h.s. of (27) satisfies,

$$\begin{aligned} \min\big\{1,\ \sqrt{\xi/L^2}\exp(-L^2/\xi)\big\} &\geq K' \frac{\sqrt{\xi}}{L}\exp(-L^2/\xi) \\ &\geq \frac{K'\sqrt{2/e}}{\sqrt{c_0}A(\lambda, n)}\frac{\sqrt{1+\lambda}}{\sqrt{\log(n)}}\exp\Big\{-\frac{e c_0 A(\lambda, n)^2}{2(1+\lambda)}\log(n)\Big\}. \end{aligned}$$

We now make explicit our choice for c_0. We pick $c_0 = 1/e$, and it can be seen that this choice is the one which asymptotically minimizes the product $c_0 A(\lambda, n)^2$. Then, the Proposition 7 below, proven in Section B.6, yields to the following bound, for a universal constant $K'' > 0$, and for n large enough,

$$\min\big\{1,\ \sqrt{\xi/L^2}\exp(-L^2/\xi)\big\} \geq \frac{K''\sqrt{1+\lambda}}{\sqrt{\log(n)}}\Big(\frac{\sqrt{\log(n)}}{n(1+\lambda)}\Big)^{\frac{e^2/2}{1+\lambda}}.$$

The last equation then shows that whenever $1 + \lambda > e^2$, the r.h.s. of (27) gets larger than the l.h.s. when n is large enough. This concludes the proof of the theorem.

Proposition 7. *Let $c_0 = 1/e$. Then whenever $1+\lambda \leq \log(n)$ we have $A(\lambda, n) = e + o(1)$ as $n \to \infty$. Furthermore when $1+\lambda \leq \log(n)$, as $n \to \infty$,*

$$c_0 A(\lambda,n)^2 \log(n) \leq e \log(n) + e \log \frac{c_3(1+\lambda)}{\sqrt{\log(n)}} + o(1).$$

4. Numerical illustrations. We present an illustration on synthetic data of the estimators introduced in Section 2. We also consider other estimators of τ_1 that have been proposed in the literature of disclosure risk assessment: i) two parametric empirical Bayes estimators of τ_1 proposed by Bethlehem et al. [2] and Skinner et al. [23]; ii) a naive nonparametric estimator of τ_1; iii) a Bayesian nonparametric estimator of τ_1 proposed by Samuels [21]. A common feature of these estimators, as well as our class of nonparametric estimators, is that they rely on modeling the random partition induced by the cross-classified sample records. More recent approaches, not considered here, focus on modeling associations among identifying variables by log-linear models, local smoothing polynomials and hierarchical latent models. In particular, the Bayesian hierarchical semiparametric models of Carota et al. [4] and Carota et al. [5] show a remarkable better performance than models for random partitions, at the cost of an increasing computational effort for the need of Markov chain Monte Carlo methods for posterior approximation.

The approach of Bethlehem et al. [2] is a parametric empirical Bayes approach in the sense of Efron and Morris [7]. It relies on the following modeling assumption for the cells' frequencies of the population: $Y_j(\boldsymbol{X}, \bar{n}) \sim \mathrm{Poiss}(\bar{n} p_j)$, where $\bar{n}$ is the size of the entire population. Bethlehem et al. [2] also assumed a Gamma prior distribution over the probabilities associated to each cell, namely $p_j \sim \mathrm{Gam}(\alpha, \beta)$. One should specify the p_j's under the condition $\sum_{j=1}^{K_{\bar{n}}} p_j = 1$, however, for the sake of simplicity, Bethlehem et al. [2] assumed that $\sum_{j=1}^{K_{\bar{n}}} \mathbb{E}[p_j] = 1$, which is tantamount to saying that $\alpha = 1/(K_{\bar{n}}\beta)$. Under these modeling assumptions, Bethlehem et al. [2] proposed an estimator of the expected value of total number $T_1(\boldsymbol{X}, \bar{n})$ of population uniques, i.e.,

$$T_1(\boldsymbol{X}, \bar{n}) := \sum_{j=1}^{K_{\bar{n}}} \mathbb{1}_{\{Y_j(\boldsymbol{X},\bar{n})=1\}}. \tag{29}$$

Under the above Poisson-Gamma model, $\mathbb{E}[T_1(\boldsymbol{X}, \bar{n})] = \bar{n}(1+\bar{n}\beta)^{-(1+\alpha)}$, which depends on the parameters α and β, with the condition $\alpha = 1/(K\beta)$. Parameters can be easily estimated via maximum likelihood, as we have done in the subsequent numerical experiments. If $K_{\bar{n}}$ is not available, Bethlehem et al. [2] suggested to estimate $K_{\bar{n}}$ assuming a uniform distribution over the cells, hence

$$\hat{K}_{\bar{n}} = \frac{\bar{n}K_n}{\sum_{j=1}^{K_n} \mathbb{1}_{\{Y_j(\boldsymbol{X},n)=1\}}},$$

where n is the size of the observed sample and K_n stands for the number of distinct cells dictated by the sample of size n. If $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators of α and β, respectively, then an estimator of $T_1(\boldsymbol{X}, \bar{n})$ is $\hat{T}_1 = \bar{n}(1+\bar{n}\hat{\beta})^{-(1+\hat{\alpha})}$. Bethlehem et al. [2] then suggested a corresponding estimator of τ_1 as the sample portion of $\hat{T}_1$. More precisely, they proposed

$$\hat{\tau}_1^B = \frac{n}{\bar{n}}\hat{T}_1 = n(1+\bar{n}\hat{\beta})^{-(1+\hat{\alpha})}. \tag{30}$$

as an estimator of τ_1. Skinner et al. [23] improved the estimator (30). In particular, still under the Poisson-Gamma model, they considered directly the problem of estimating τ_1. In particular, they proposed the following estimator

$$\hat{\tau}_1^S := K_n \left(\frac{1+\bar{n}\hat{\beta}}{1+n\hat{\beta}}\right)^{-(1+\hat{\alpha})}, \tag{31}$$

where the prior parameters α and β can be estimated via maximum likelihood. The estimators proposed in Section 2, due to their nonparametric empirical Bayes interpretation in the sense of Robbins [20], may be considered as the natural nonparametric counterparts of the empirical Bayes estimator (31).

Besides parametric estimators of τ_1, we also consider two nonparametric estimators. A naive nonparametric estimator of τ_1 relies on the intuition that a natural estimator of τ_1 is the sampling fraction, with respect to the population, of the number of sample uniques. This estimator was first discussed in Bethlehem et al. [2] and Skinner et al. [23], and it is defined as follows

$$\hat{\tau}_1^{\mathscr{N}} := Z_1(\boldsymbol{X}, n)\frac{n}{\bar{n}}. \tag{32}$$

Samuels [21] exploits Bayesian nonparametric ideas, and in particular a Dirichlet process prior (Ferguson [9]) on the p_j's to derive a smoothed version of the naive estimator (32). In particular, Samuels [21] suggested the following estimator

$$\hat{\tau}_1^{\mathscr{D}} := Z_1(\boldsymbol{X}, n)\frac{n+\vartheta-1}{\bar{n}+\vartheta-1}, \tag{33}$$

where ϑ is the concentration parameter of the Dirichlet process prior. It is well-known (see, e.g. Ferguson [9]) that the maximum likelihood estimator of ϑ can be obtained by solving, with respect to ϑ, the equation $K_n = \sum_{1\leq j\leq n-1}\vartheta/(\vartheta+j)$.

We generate synthetic tables with C cells, where $C = 3\cdot 10^5$ for Table 1, $C = 6\cdot 10^5$ for Table 2 and $C = 9\cdot 10^5$ for Table 3. For any choice of the size C, the true probabilities $(p_j)_{j\geq 1}$ of cells have been generated according to different types of distributions: the Zipf distribution, i.e., $p_j \propto j^{-s}$ for some $s > 0$, the uniform distribution over the total number of cells and the uniform Dirichlet distribution. For all the simulated scenarios we have considered a population of size $\bar{n} = 10^6$ and a sample of $n = 10^5$ individuals from it. Each column of Tables 1–3 corresponds to a different choice of the distribution over the cells' probabilities. From the left to right: the Zipf distribution with parameter $s = 0.2, 0.5, 0.8, 1$, the uniform distribution, the uniform Dirichlet distribution with parameter $\beta = 0.5, 1$. In the first row of each table we have reported the true values of the disclosure index, while the other rows contain the estimates obtained with: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$, see Proposition 2; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$, see Proposition 1; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^{B}$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^{S}$ All experiments are averaged over 100 iterations. The best estimates in each simulated scenarios are displayed in bold.

	Zipf 0.2	Zipf 0.5	Zipf 0.8	Zipf 1	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	2868	4651	7537	7313	2579	4151	4608
$\hat{\tau}_1^{L_b}$	**5671**	9322	12790	10373	**5582**	2814	3118
$\hat{\tau}_1^{L_p}$	21086	20451	18221	12744	21656	8205	11855
$\hat{\tau}_1^{\mathscr{N}}$	6413	**5613**	3810	2157	6511	**4232**	**5111**
$\hat{\tau}_1^{\mathscr{D}}$	18461	12471	**5421**	2459	19303	7498	10741
$\hat{\tau}_1^{B}$	30554	27271	13848	**5358**	30847	22820	26351
$\hat{\tau}_1^{S}$	28702	23621	9670	3043	29187	17665	22306

TABLE 1
Estimators of τ_1 for several simulated scenarios, when the size of the table is $C = 3\cdot 10^5$.

	Zipf 0.2	Zipf 0.5	Zipf 0.8	Zipf 1	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	16020	16819	16095	11401	15947	9857	12451
$\hat{\tau}_1^{L_b}$	32254	27585	21216	**13194**	33426	**9190**	**14805**
$\hat{\tau}_1^{L_p}$	49883	42310	29344	16847	51096	22932	32146
$\hat{\tau}_1^{\mathscr{N}}$	**7625**	6860	4642	2534	**7693**	5899	6670
$\hat{\tau}_1^{\mathscr{D}}$	32567	**21009**	7380	2968	33860	14577	20337
$\hat{\tau}_1^{B}$	33594	31155	**17152**	6380	33763	28795	31114
$\hat{\tau}_1^{S}$	34070	29703	13032	3776	34321	25900	29634

Table 2

Estimators of τ_1 for several simulated scenarios, when the size of the table is $C = 6 \cdot 10^5$.

	Zipf 0.2	Zipf 0.5	Zipf 0.8	Zipf 1	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	28976	27049	21933	13794	29483	15635	20281
$\hat{\tau}_1^{L_b}$	49619	41086	28366	**15980**	51607	**17952**	29496
$\hat{\tau}_1^{L_p}$	63328	53867	35646	19764	64964	34447	46199
$\hat{\tau}_1^{\mathscr{N}}$	8076	7406	5082	2729	8138	6729	7371
$\hat{\tau}_1^{\mathscr{D}}$	42337	**27383**	8651	3246	44104	20789	**28307**
$\hat{\tau}_1^{B}$	**34628**	32675	**18977**	6942	**34772**	31235	32935
$\hat{\tau}_1^{S}$	35957	32338	15028	4196	36234	29837	32804

Table 3

Estimators of τ_1 for several simulated scenarios, when the size of the table is $C = 9 \cdot 10^5$.

From Tables 1–3, we observe that the choice of the smoothing distribution L for $\hat{\tau}_1^{L}$, i.e. the Binomial smoothing or the Poisson smoothing, is crucial with respect to the performance of the corresponding estimators. In particular, in all the simulated scenarios the Binomial smoothing displays a better performance than the Poisson smoothing. We also observe that the performance of the estimators strongly depends on the size of the contingency table along with the distributions over the cell's probabilities. However, there is not a clear path indicating in which simulated scenarios nonparametric estimators outperform parametric estimators. In general, we may say that nonparametric estimators have a better performance than parametric estimators when the size of the contingency table is relative small. This confirm a phenomenon already observed in the experimental analysis presented in Samuels [21] for $\hat{\tau}_1^{\mathscr{D}}$. In general, the nonparametric estimator $\hat{\tau}_1^{L_b}$ has a good performance when the cells probabilities follow the uniform Dirichlet distribution, the uniform distribution and the Zipf with parameter 1. Also, $\hat{\tau}_1^{L_b}$ appears to be more accurate when the size of the contingency table is relative small.

APPENDIX A: NONPARAMETRIC ESTIMATORS OF THE DISCLOSURE RISK: PROOFS

Here we will prove all the results stated in Section 2. For the sake of simplifying notations, we will simply write τ_1 instead of $\tau_1(\boldsymbol{X}, N, M)$, as well as $\hat{\tau}_1$ (resp. $\hat{\tau}_1^L$) instead of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ (resp. $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$).

A.1. Details for the determination of (2). First of all observe that the expected value of $Z_i(\boldsymbol{X}, N)$ can be easily computed

$$\begin{aligned}\mathbb{E}[Z_i(\boldsymbol{X}, N)] &= \mathbb{E}\left[\sum_{j\geq 1} \mathbb{1}_{\{Y_j(\mathbf{X},N)=i\}}\right] = \sum_{j\geq 1} \mathbb{P}(Y_j(\mathbf{X}, N) = i) \\ &= \sum_{j\geq 1} e^{-np_j} \frac{(np_j)^i}{i!}.\end{aligned} \tag{A.1}$$

Using the Taylor series expansion of the exponential function, we get

$$\begin{aligned}\mathbb{E}[\tau_1] &= \mathbb{E}\left[\sum_{j\geq 1} \mathbb{1}_{\{Y_j(\mathbf{X},N)=1\}} \mathbb{1}_{\{Y_j(\mathbf{X},N+M)=1\}}\right] \\ &= \sum_{j\geq 1} \mathbb{P}(Y_j(\mathbf{X}, N) = 1)\mathbb{P}(Y_j(\mathbf{X}, N+M) - Y_j(\mathbf{X}, N) = 0) \\ &= \sum_{j\geq 1} np_j e^{-np_j} e^{-\lambda np_j} = \sum_{j\geq 1} np_j e^{-np_j} \sum_{i\geq 0} \frac{(-\lambda np_j)^i}{i!} \\ &= \sum_{i\geq 0} \frac{(-1)^i \lambda^i}{i!} \sum_{j\geq 1} (np_j)^{i+1} e^{-np_j} = \sum_{i\geq 0} (-1)^i \lambda^i (i+1)\mathbb{E}[Z_{i+1}(\boldsymbol{X}, N)]\end{aligned}$$

where the last equality follows from (A.1).

A.2. Empirical Bayes approach to determine (3). The estimator $\hat{\tau}_1$ can be derived as the empirical Bayes estimator of $\mathbb{E}[\hat{\tau}_1]$ in the sense of [20], see also [15] for an overview of the empirical Bayes approach. First, it is worth noticing that the expectation of τ_1 coincides with

$$\mathbb{E}[\tau_1] = \sum_{j=1}^{+\infty} e^{-(\lambda+1)np_j} np_j. \tag{A.2}$$

We observe that the statistic τ_1 is a function of the observations only through the frequency counts $Y_j(\boldsymbol{X}, N)$, which, in our model, are Poisson distributed with parameter np_j. In order to derive the nonparametric empirical Bayes estimator of (A.2), we assume that $p_1, p_2, \ldots$ are independent and distributed according to the empirical cumulative distribution function $G(p)$ of $p_{i_1}, \ldots, p_{i_k}$, corresponding to the k distinct cells arising from the cross classification of the initial sample, namely $G(p) := k^{-1} \sum_{1 \le t \le k} \mathbb{1}_{\{p_{i_t} \le p\}}$. Consider a cell j containing x individuals out of the initial sample of size N, where $x \ge 0$, then by a proper adaptation of [20, formula (9)] to our setting, we find out that

$$\varphi_n(x) := \frac{\int e^{-(\lambda+1)np} npe^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \tag{A.3}$$

is the Bayes estimator of the quantity $e^{-(\lambda+1)np_j} np_j$ appearing in (A.2) for a cell j which contains x individuals out of the initial sample of size N. We now rewrite $\varphi_n(x)$ as follows

$$\begin{aligned}
\varphi_n(x) &= \frac{\int e^{-(\lambda+1)np} npe^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\
&= \frac{\int \sum_{i \ge 0} \frac{(-(\lambda+1)np)^i}{i!} npe^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\
&= \frac{\sum_{i \ge 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \int \frac{(np)^{x+i+1}}{(x+i+1)!} e^{-np} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\
&= \frac{\sum_{i \ge 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \mathbb{E}[Z_{x+i+1}(\boldsymbol{X}, N)]}{\mathbb{E}[Z_x(\boldsymbol{X}, N)]}.
\end{aligned}$$

Then the nonparametric Bayes estimator of $\mathbb{E}[\tau_1]$ may be obtained summing up over all the possible cross classification of the observed cells:

$$\begin{aligned}
&\sum_{x \ge 0} Z_x(\boldsymbol{X}, N) \varphi_n(x) \\
&\qquad = \sum_{x \ge 0} Z_x(\boldsymbol{X}, N) \frac{\sum_{i \ge 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \mathbb{E}[Z_{x+i+1}(\boldsymbol{X}, N)]}{\mathbb{E}[Z_x(\boldsymbol{X}, N)]}
\end{aligned} \tag{A.4}$$

The empirical Bayes estimator of $\mathbb{E}[\tau_1]$ coincides with (A.4) where we replace the expectations $\mathbb{E}[Z_x(\boldsymbol{X}, N)]$ with their empirical counterparts $Z_x(\boldsymbol{X}, N)$:

$$\hat{\tau}_1 = \sum_{x \ge 0} Z_x(\boldsymbol{X}, N) \frac{\sum_{i \ge 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! Z_{x+i+1}(\boldsymbol{X}, N)}{Z_x(\boldsymbol{X}, N)}$$

$$
\begin{aligned}
&= \sum_{x\geq 0}\sum_{i\geq 0} \frac{(-(\lambda+1))^i}{i!x!}(x+i+1)!Z_{x+i+1}(\boldsymbol{X},N) \\
&= \sum_{x\geq 0}\sum_{i\geq x} \frac{(-(\lambda+1))^{i-x}}{(i-x)!x!}(i+1)!Z_{i+1}(\boldsymbol{X},N) \\
&= \sum_{i\geq 0}(i+1)Z_{i+1}(\boldsymbol{X},N)\sum_{x=0}^{i}\frac{i!}{(i-x)!x!}(-(\lambda+1))^{i-x} \\
&= \sum_{i\geq 0}(-1)^i\lambda^i(i+1)Z_{i+1}(\boldsymbol{X},N),
\end{aligned}
$$

hence (3) now follows.

A.3. Proof of Theorem 1. The unbiasedness of $\hat{\tau}_1$ follows from (2). Hence we focus on the proof of the variance bound (5). Thanks to the independence of the random variables $\{Y_j(\boldsymbol{X},N)\}_{j\geq 1}$, we may write the variance $\mathrm{Var}(\tau_1-\hat{\tau}_1)$ as

$$
\sum_{j\geq 1}\mathrm{Var}\left(\sum_{i\geq 0}(-1)^i(i+1)\lambda^i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right).
$$

Now the unbiasedness of the estimator implies

$$
\begin{aligned}
&\mathrm{Var}(\tau_1-\hat{\tau}_1) \\
&\quad= \sum_{j\geq 1}\mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right]^2 \\
&\quad= \sum_{j\geq 1}\mathbb{E}\left[\sum_{i\geq 1}a_i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\left(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)\right]^2,
\end{aligned}
$$

where we have defined

$$
a_i := (-1)^i(i+1)\lambda^i.
$$

It is now easy to observe that the events $\{(Y_j(\boldsymbol{X},N)=i)\}_{i\geq 1}$ are all disjoint, hence the variance $\mathrm{Var}(\tau_1-\hat{\tau}_1)$ may be rewritten as

$$
\sum_{j\geq 1}\mathbb{E}\left[\sum_{i\geq 1}a_i^2\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\left(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)^2\right]
$$

$$= \sum_{j\geq 1} \mathbb{E}\left[\sum_{i\geq 0} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right]$$

observing that $a_0 = 1$. Thus, simple calculations show that we can bound the variance as follows

$$\begin{aligned}
\mathrm{Var}(\tau_1 - \hat{\tau}_1) &\leq \max_{j\geq 0} |a_j|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X},N)] - \sum_{j\geq 1} e^{-n(\lambda+1)p_j} np_j \\
&= \max_{i\geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X},N)] - \frac{1}{\lambda+1}\mathbb{E}[Z_1(\boldsymbol{X},N+M)]. \qquad \text{(A.5)}
\end{aligned}$$

To conclude the proof, it remains to show that the a_i's have a maximum for $\lambda < 1$, which is attained when $i = i^* := \lfloor (2\lambda-1)/(1-\lambda)\rfloor \vee 0$. Hence the thesis follows by (A.5), realizing that $\max_{i\geq 0}|a_i| = \Psi(\lambda)$.

A.4. Proof of Theorem 2. First we focus on the determination of the bound (8), concerning the bias. Remember the definition of both $\hat{\tau}_1^L$ and τ_1 to write

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_1^L - \tau_1] &= \mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L\geq i) Z_{i+1}(\boldsymbol{X},N)\right.\\
&\qquad\qquad \left. - \sum_{j\geq 1}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right]\\
&= -\mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L\leq i-1) Z_{i+1}(\boldsymbol{X},N)\right]
\end{aligned}$$

where we have observed that non–smoothed estimator $\hat{\tau}_1$ is unbiased. It is now easy to see that

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_1^L - \tau_1] &= -\mathbb{E}\left[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L\leq i-1) Z_{i+1}(\boldsymbol{X},N)\right]\\
&= -\mathbb{E}\left[\sum_{i\geq 1}(-1)^i(i+1)\lambda^i \mathbb{P}(L\leq i-1) \sum_{j\geq 1}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}}\right]\\
&= -\sum_{i\geq 1}\sum_{j\geq 1}(-1)^i(i+1)\lambda^i \mathbb{P}(L\leq i-1)\mathbb{P}(Y_j(\boldsymbol{X},N)=i+1)\\
&= -\sum_{i\geq 1}\sum_{j\geq 1}(-1)^i(i+1)\lambda^i \mathbb{P}(L\leq i-1) e^{-np_j}\frac{(np_j)^{i+1}}{(i+1)!}
\end{aligned}$$

$$= -\sum_{j\geq 1} e^{-np_j} np_j \sum_{i\geq 1} (-1)^i \frac{(\lambda np_j)^i}{i!} \mathbb{P}(L \leq i-1). \tag{A.6}$$

Now we focus on the evaluation of the sum with respect to i, for the sake of clarity we write $y := \lambda np_j$, hence

$$\begin{aligned}\sum_{i\geq 1}(-1)^i \frac{y^i}{i!}\mathbb{P}(L\leq i-1) &= \sum_{i=1}^{+\infty}(-1)^i\frac{y^i}{i!}\sum_{k=0}^{i-1}\mathbb{P}(L=k)\\ &= \sum_{k=0}^{+\infty}\mathbb{P}(L=k)\sum_{i=k+1}^{+\infty}\frac{(-y)^i}{i!}\end{aligned}$$

and remembering the definition of the incomplete gamma function we obtain that

$$\begin{aligned}\sum_{i\geq 1}(-1)^i\frac{y^i}{i!}\mathbb{P}(L\leq i-1) &= \sum_{k=0}^{+\infty}\mathbb{P}(L=k)\frac{e^{-y}}{k!}\int_0^{-y}\tau^k e^{-\tau}\mathrm{d}\tau\\ &= -\sum_{k=0}^{+\infty}\mathbb{P}(L=k)\frac{e^{-y}}{k!}\int_0^{y}(-s)^k e^{s}\mathrm{d}s\\ &= -e^{-y}\int_0^y e^s\mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]\mathrm{d}s.\end{aligned}$$

Putting the previous expression in (A.6) and observing that $y = \lambda np_j$, (8) immediately follows.

We are ready to bound the variance of the difference between τ_1 and its estimator $\hat{\tau}_1^L$. Recalling that the random variables $\{Y_j(\boldsymbol{X}, N)\}_{j\geq 1}$ are independent, a direct calculation shows that

$$\begin{aligned}\mathrm{Var}(\hat{\tau}_1^L - \tau_1) &= \mathrm{Var}\Bigg(\sum_{i\geq 0}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X},N)\mathbb{P}(L\geq i)\\ &\qquad -\sum_{j=1}^{+\infty}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg)\\ &= \sum_{j=1}^{+\infty}\mathrm{Var}\Big(\sum_{i=0}^{+\infty}(-1)^i(i+1)\lambda^i\mathbb{P}(L\geq i)\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}}\\ &\qquad -\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Big)\\ &= \sum_{j=1}^{+\infty}\mathrm{Var}\left(\sum_{i=0}^{+\infty}a_i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right),\end{aligned}$$

having defined

$$a_i := (-1)^i (i+1)\lambda^i \mathbb{P}(L \geq i)$$

for any $i \geq 0$. Hence the variance of $\hat{\tau}_1^L - \tau_1$ may be upper bounded by the quantity

$$\begin{aligned}
&\sum_{j=1}^{+\infty} \mathbb{E}\left[\left(\sum_{i=0}^{+\infty} a_i \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right)^2\right] \\
&\quad = \sum_{j=1}^{+\infty} \mathbb{E}\left[\left(\sum_{i=1}^{+\infty} a_i \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} (a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}})\right)^2\right] \\
&\quad = \sum_{j=1}^{+\infty} \mathbb{E}\left[\sum_{i=1}^{+\infty} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} (a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}})^2\right]
\end{aligned}$$

where we have used the incompatibility of the events $\{(Y_j(\boldsymbol{X}, N) = i)\}$ for different values of j. We can proceed with the upper bound for the variance as follows

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}_1^L - \tau_1) &= \sum_{j=1}^{+\infty} \mathbb{E}\left[\sum_{i=0}^{+\infty} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right] \\
&\leq \max_{i \geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \sum_{j=1}^{+\infty} \mathbb{E}\left[\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\right] \\
&= \max_{i \geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \sum_{j=1}^{+\infty} e^{-\lambda n p_j} e^{-n p_j} n p_j \\
&= \max_{i \geq 0} |a_i|^2 \mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{1}{\lambda+1} \mathbb{E}[Z_1(\boldsymbol{X}, N+M)]. \qquad \text{(A.7)}
\end{aligned}$$

We can estimate the maximum value of the $|a_i|$'s as follows

$$\begin{aligned}
\max_{i \geq 0} |a_i| &= \max_{i \geq 0} (i+1)\lambda^i \mathbb{P}(L \geq i) = \max_{i \geq 0} (i+1)\lambda^i \sum_{k=i}^{+\infty} \mathbb{P}(L = k) \\
&\leq \max_{i \geq 0} \sum_{k=i}^{+\infty} (i+1)\lambda^i \mathbb{P}(L = k) \leq \sum_{k=0}^{+\infty} (k+1)\lambda^k \mathbb{P}(L = k) \\
&= \mathbb{E}_L[(L+1)\lambda^L].
\end{aligned}$$

Hence, replacing $\max_{i\geq 0}|a_i|$ with $\mathbb{E}_L[(L+1)\lambda^L]$ in (A.7), the upper bound for the variance becomes

$$\text{(A.8)}\quad \mathrm{Var}(\hat{\tau}_1^L - \tau_1) \leq (\mathbb{E}_L[(L+1)\lambda^L])^2\mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1}.$$

Putting together the bound for the variance (A.8) and for the bias (8), the bound on the MSE (9) easily follows.

A.5. Proof of Proposition 1. Let us now prove the bound (10) on the MSE, in order to do this, we use Theorem 2, bounding the two terms appearing in (9) separately.
To obtain an estimate of first term on the r.h.s. of (9), we note that for any $y > 0$ the following holds

$$\begin{aligned}-e^{-y}\int_0^y e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]\mathrm{d}s &= -e^{-y}\int_0^y e^s \sum_{k=0}^{+\infty} e^{-\beta}\frac{\beta^k}{k!}\frac{(-s)^k}{k!}\mathrm{d}s\\ &= -e^{-y-\beta}\int_0^y e^s \sum_{k=0}^{+\infty}\frac{(\beta s)^k(-1)^k}{\Gamma(k+1)k!}\mathrm{d}s\end{aligned}$$

Recall that the Bessel polynomial (see Olver et al. [1]) is defined as

$$J_0(z) := \sum_{k=0}^{+\infty}\frac{(-1)^k z^{2k}}{2^{2k}\Gamma(k+1)k!}.$$

and that $|J_0(z)| \leq 1$, hence we obtain

$$\begin{aligned}&\left|-e^{-y}\int_0^y e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]\mathrm{d}s\right|\\ &\qquad\leq e^{-(y+\beta)}\int_0^y e^s |J_0(2\sqrt{s\beta})|\mathrm{d}s \leq e^{-\beta}(1-e^{-y}).\end{aligned}$$

The previous estimate may be applied to bound the first term on the r.h.s. of (9), with $y = \lambda n p_j$, indeed

$$\text{(A.9)}\quad \begin{aligned}&\left|\sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]\mathrm{d}s\right|\\ &\qquad\leq \sum_{j\geq 1} e^{-np_j} np_j e^{-\beta}(1-e^{-\lambda np_j}) \leq e^{-\beta}\sum_{j=1}^{+\infty} e^{-np_j} np_j\\ &\qquad= e^{-\beta}\mathbb{E}[Z_1(\boldsymbol{X}, N)] \leq e^{-\beta}\mathbb{E}[N] = e^{-\beta} n.\end{aligned}$$

having observed that the maximum number of species with frequency one in a sample of size N is exactly N.
Second, in order to upper bound the other term on the r.h.s of (9), we observe that

$$\begin{aligned}\mathbb{E}_L[(L+1)\lambda^L] &= \sum_{k=0}^{+\infty} e^{-\beta}\frac{\beta^k}{k!}\lambda^k(k+1)\\ &= e^{-\beta}\left(\sum_{k=1}^{+\infty}\frac{(\beta\lambda)^k}{(k-1)!}+\sum_{k=0}^{+\infty}\frac{(\beta\lambda)^k}{k!}\right)\\ &= e^{-\beta}(e^{\beta\lambda}+\beta\lambda e^{\beta\lambda}) = e^{\beta(\lambda-1)}(1+\beta\lambda),\end{aligned}$$

hence we get

$$\begin{aligned}(\mathbb{E}_L[(L+1)\lambda^L])^2\mathbb{E}[Z_{\bar{1}}(\boldsymbol{X},N)] - \frac{1}{\lambda+1}\mathbb{E}[Z_1(\boldsymbol{X},N+M)]&\\ \leq ne^{2\beta(\lambda-1)}(1+\beta\lambda)^2.&\end{aligned} \tag{A.10}$$

Using (A.9) and (A.10), one can now estimate the MSE (9) in the Poisson case and (10) follows.

Thanks to (10) just derived, the normalized mean square error can be bounded from above by

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq e^{-2\beta} + \frac{e^{2\beta(\lambda-1)}(1+\beta\lambda)^2}{n}$$

using the exponential inequality $1+x\leq e^x$ we get

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq e^{-2\beta} + \frac{e^{2\beta(2\lambda-1)}}{n}. \tag{A.11}$$

The r.h.s. of (A.11) is minimized when β equals $\frac{1}{4\lambda}\log\left(\frac{n}{2\lambda-1}\right)$, it is easy to observe that (A.11) becomes

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \frac{1}{n^{1/(2\lambda)}}\cdot\frac{2\lambda}{(2\lambda-1)^{1-1/(2\lambda)}} \tag{A.12}$$

hence the second bound (11) follows provided that

$$A(\lambda) := \cdot\frac{2\lambda}{(2\lambda-1)^{1-1/(2\lambda)}}.$$

We are now ready to prove the limit of predictability in the Poisson case, indeed thanks to (11) we have

$$\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \frac{A}{n^{1/(2\lambda)}},$$

besides observe that the inequality

$$\frac{A}{n^{1/(2\lambda)}} \leq \delta$$

is satisfied iff

$$\lambda \leq \frac{\log(n)}{2\log(A/\delta)} =: \lambda^*.$$

As a consequence the maximum value of λ for which the inequality $\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta$ is satisfied, is bigger or equal than λ^*, in other words

$$\max\left\{\lambda : \ \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta\right\} \geq \frac{\log(n)}{2\log(A/\delta)}.$$

The thesis follows by taking the limit of the previous inequality as $n \to +\infty$.

APPENDIX B: PROOFS OF AUXILIARY RESULTS FOR THE LOWER BOUND

B.1. Proof of Lemma 1. First, it is obvious that

$$\mathscr{E}(\lambda, n) \leq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2].$$

We now prove that the previous is indeed an inequality by deriving a lower bound that essentially matches. Let $n > 0$ be fixed. By definition, for every $\varepsilon > 0$ there exists an estimator $\hat{\rho}_1$ such that

$$\begin{aligned}
\mathscr{E}(\lambda, n) &\geq \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(\boldsymbol{X}(N), N))^2] - \varepsilon \\
&= \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[\mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) \\
&\qquad\qquad - \hat{\rho}_1(\boldsymbol{X}(N), N))^2 \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N+M)]] - \varepsilon \\
&\geq \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \mathbb{E}_P^n[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)])^2] - \varepsilon \qquad \text{(B.1)}
\end{aligned}$$

where the last line follows by Jensen's inequality and by observing that

$$\mathbb{E}_P^n[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N+M)] = \tau_1(\boldsymbol{X}, N, M), \quad \text{and,}$$
$$\mathbb{E}_P^n[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N+M)] = \mathbb{E}_P^n[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)].$$

To see that the last equation is true, remark that $\boldsymbol{Y}(\boldsymbol{X}, N+M)-\boldsymbol{Y}(\boldsymbol{X}, N)$ is independent of $\boldsymbol{Y}(\boldsymbol{X}, N)$ and depends only on $(X_{N+1}, \dots, X_{N+M})$. Now we claim that $\hat{\rho}_1$ can be chosen such that for any $k \in \mathbb{Z}_+$ and any permutation $\sigma_k(\boldsymbol{X}(k))$ of the data, it holds $\hat{\rho}_1(\boldsymbol{X}(k), k) = \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k)$. We delay the proof of the claim to later. Now assume the claim is true. Given k and $\boldsymbol{Y}(\boldsymbol{X}, k)$, we can construct the functional

$$G(\boldsymbol{Y}(\boldsymbol{X}, k), k) := (\underbrace{1, \dots, 1}_{\times Y_1(\boldsymbol{X}, k)}, \underbrace{2, \dots, 2}_{\times Y_2(\boldsymbol{X}, k)}, \dots).$$

Since $\hat{\rho}_1$ is invariant under permutations of the data, we have for any $P \in \mathscr{P}$,

$$\begin{aligned}
&\mathbb{E}_P^n[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] \\
&\qquad = \mathbb{E}_P^n\big[\mathbb{E}_P^n[\hat{\rho}_1(\boldsymbol{X}(N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N), N] \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\qquad = \mathbb{E}_P^n\big[\mathbb{E}_P^n[\hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N), N] \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\qquad = \mathbb{E}_P^n[\hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] \\
&\qquad = \hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N).
\end{aligned}$$

The last line follows because $N = \sum_{j \geq 1} Y_j(\boldsymbol{X}, N)$, and hence N is completely determined by $\boldsymbol{Y}(\boldsymbol{X}, N)$. Therefore, we have proved that the conditional expected value of $\hat{\rho}_1(\boldsymbol{X}(N), N)$, given $\boldsymbol{Y}(\boldsymbol{X}, N)$ does not depend on P. Thus, (B.1) implies,

$$\begin{aligned}
\mathscr{E}(\lambda, n) &\geq \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1(G(\boldsymbol{Y}(\boldsymbol{X}, N), N), N))^2] - \varepsilon \\
&\geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^n[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2] - \varepsilon.
\end{aligned}$$

Since the previous is true for all $\varepsilon > 0$, the conclusion follows.

We now prove the claim we have used in the previous argument, i.e. that $\hat{\rho}_1$ *can be chosen such for any* $k \in \mathbb{Z}_+$ *and any permutation* $\sigma_k(\boldsymbol{X}(k))$ *of the data, it holds* $\hat{\rho}_1(\boldsymbol{X}(k), k) = \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k)$. When $k = 0$, then the claim is trivial, hence we assume without loss of generality that $k \in \mathbb{N}$. We will prove that for any estimator $\hat{\rho}_1$, there is a symmetric estimator $\hat{t}_1$ with a risk no more than the risk of $\hat{\rho}_1$. Let $\hat{\rho}_1$ be arbitrary. Construct $\hat{t}_1$ such that for any $k \in \mathbb{N}$

$$\hat{t}_1(\boldsymbol{X}(k), k) := \frac{1}{|\{\sigma_k\}|} \sum_{\{\sigma_k\}} \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k).$$

Clearly $\hat{t}_1$ has the desired invariance property under permutations. Moreover, by Jensen's inequality,

$$
\begin{aligned}
&\mathbb{E}_P^n[(\tau_1(\boldsymbol{X},N,M) - \hat{t}_1(\boldsymbol{X}(N),N))^2] \\
&\quad = \mathbb{E}_P^n\Big[\mathbb{E}_P^n\Big[\Big(\frac{1}{|\{\sigma_N\}|}\sum_{\{\sigma_N\}}(\tau_1(\boldsymbol{X},N,M) - \hat{\rho}_1(\sigma_N(\boldsymbol{X}(N)),N)\Big)^2 \mid N\Big]\Big] \\
&\quad \leq \mathbb{E}_P^n\Big[\mathbb{E}_P^n\Big[\frac{1}{|\{\sigma_N\}|}\sum_{\{\sigma_N\}}(\tau_1(\boldsymbol{X},N,M) - \hat{\rho}_1(\sigma_N(\boldsymbol{X}(N)),N))^2 \mid N\Big]\Big]
\end{aligned}
$$

Now remark that for all $(k,k') \in \mathbb{Z}_+^2$ the map $\boldsymbol{X} \mapsto \tau_1(\boldsymbol{X},k,k')$ is invariant under any permutations of the k first entries of $\boldsymbol{X}$. Moreover, $\boldsymbol{X}$ is an i.i.d. vector, then the last display implies that

$$
\begin{aligned}
&\mathbb{E}_P^n[(\tau_1(\boldsymbol{X},N,M) - \hat{t}_1(\boldsymbol{X}(N),N))^2] \\
&\qquad = \mathbb{E}_P^n\Big[\mathbb{E}_P^n\Big[\frac{1}{|\{\sigma_N\}|}\sum_{\{\sigma_N\}}(\tau_1(\boldsymbol{X},N,M) - \hat{\rho}_1(\boldsymbol{X}(N),N))^2 \mid N\Big]\Big] \\
&\qquad = \mathbb{E}_P^n[(\tau_1(\boldsymbol{X},N,M) - \hat{\rho}_1(\boldsymbol{X}(N),N))^2].
\end{aligned}
$$

The conclusion follows by taking the supremum over $P \in \mathscr{P}$ both sides of the last display.

B.2. Proof of Proposition 3. For any $P \in \mathscr{P}'$ we write $\tilde{P}(\cdot) := P(\cdot)/P(\mathbb{N})$, so that $\tilde{P} \in \mathscr{P}$ is a probability measure. We write $\tilde{p}_j := p_j/P(\mathbb{N})$, $j \in \{1,\dots,S\}$. Furthermore we let $m(P) := n\sum_{j=1}^S p_j$. Then since $\boldsymbol{Y}$ is a vector of independent Poisson random variables, is clear that for any $P \in \mathscr{P}'$

$$
\mathbb{E}_{\tilde{P}}^n[(\bar{\tau}_1(\tilde{P},n) - \hat{\rho}(\boldsymbol{Y}))^2] = \mathbb{E}_P^{m(P)}[(\bar{\tau}_1(\tilde{P},n) - \hat{\rho}(\boldsymbol{Y}))^2]. \tag{B.2}
$$

We now choose $\hat{\tau}$ to be an estimator satisfying for some $\zeta > 0$

$$
\sup_{P\in\mathscr{P}'} \mathbb{E}_P^{m(P)}[(\bar{\tau}_1(\tilde{P},n) - \hat{\tau}(\boldsymbol{Y}))^2] \leq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}'} \mathbb{E}_P^{m(P)}[(\bar{\tau}_1(\tilde{P},n) - \hat{\rho}(\boldsymbol{Y}))^2] + \zeta.
$$

This is always possible for any $\zeta > 0$. Furthermore remark that $m(P) \leq (1 + c_1\varepsilon/\xi)n = n'$, so that $m(P)/n' \leq 1$ always when $P \in \mathscr{P}'$. Let $P \in \mathscr{P}'$ be fixed, and let $\boldsymbol{W} = (W_1, W_2, \dots)$ such that conditional on $\boldsymbol{Y}$, the random variables W_j are independent binomial random variables with parameters $(Y_j, m(P)/n')$. Then define $\tilde{\tau}(\boldsymbol{Y}) \coloneqq \mathbb{E}[\hat{\tau}(\boldsymbol{W})] \mid \boldsymbol{Y}]$. By Jensen's inequality,

$$
\begin{aligned}
\mathbb{E}_P^{n'}[(\bar{\tau}_1(\tilde{P},n) - \tilde{\tau}(\boldsymbol{Y}))^2] &= \mathbb{E}_P^{n'}[(\mathbb{E}[\bar{\tau}_1(\tilde{P},n) - \hat{\tau}(\boldsymbol{W}) \mid \boldsymbol{Y}])^2] \\
&\leq \mathbb{E}_P^{n'}[\mathbb{E}[(\bar{\tau}_1(\tilde{P},n) - \hat{\tau}(\boldsymbol{W}))^2 \mid \boldsymbol{Y}]] \\
&= \mathbb{E}_P^{m(P)}[(\bar{\tau}_1(\tilde{P},n) - \hat{\tau}(\boldsymbol{Y}))^2]
\end{aligned}
$$

$$\leq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_P^{m(P)}[(\bar{\tau}_1(\tilde{P}, n) - \hat{\rho}(\boldsymbol{Y}))^2] + \zeta.$$

Taking the supremum over $P \in \mathscr{P}'$ on the lhs of the last display, and using that the infinimum over $\hat{\rho}$ will be always smaller than the value at $\tilde{\tau}$, we find using (B.2) that

$$\begin{aligned}\inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} \mathbb{E}_P^n[(\bar{\tau}_1(P, n) - \hat{\rho}(\boldsymbol{Y}))^2] &= \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{\tilde{P}}^n[(\bar{\tau}_1(\tilde{P}, n) - \hat{\rho}(\boldsymbol{Y}))^2] \\ &= \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_P^{m(P)}[(\bar{\tau}_1(\tilde{P}, n) - \hat{\rho}(\boldsymbol{Y}))^2] \\ &\geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_P^{n'}[(\bar{\tau}_1(\tilde{P}, n) - \hat{\rho}(\boldsymbol{Y}))^2] - \zeta\end{aligned}$$

Since the previous is true for all $\zeta > 0$, we indeed have proven

$$\inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} \mathbb{E}_P^n[(\bar{\tau}_1(P, n) - \hat{\rho}(\boldsymbol{Y}))^2] \geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_P^{n'}[(\bar{\tau}_1(\tilde{P}, n) - \hat{\rho}(\boldsymbol{Y}))^2]. \tag{B.3}$$

To finish the proof of the proposition, we will now show that $\bar{\tau}_1(\tilde{P}, n)$ in (B.3) can be traded for $\bar{\tau}_1(P, n)$ at small cost. Remark that by Young's inequality, for any $P \in \mathscr{P}'$ and any $\hat{\rho}$,

$$\begin{aligned}&\mathbb{E}_P^{n'}[(\bar{\tau}_1(\tilde{P}, n) - \hat{\rho}(\boldsymbol{Y}))^2] \\ &\qquad \geq \frac{1}{2}\mathbb{E}_P^{n'}[(\bar{\tau}_1(P, n) - \hat{\rho}(\boldsymbol{Y}))^2] - (\bar{\tau}_1(P, n) - \bar{\tau}_1(\tilde{P}, n))^2,\end{aligned} \tag{B.4}$$

with

$$\begin{aligned}&\bar{\tau}_1(P, n) - \bar{\tau}_1(\tilde{P}, n) \\ &\qquad = -n\sum_{j=1}^S (\tilde{p}_j - p_j)e^{-(1+\lambda)np_j} + n\sum_{j=1}^S \tilde{p}_j e^{-(1+\lambda)np_j}\Big\{1 - e^{n(1+\lambda)(p_j - \tilde{p}_j)}\Big\}.\end{aligned}$$

Thanks to a Taylor expansion of the term within the brackets in the last display for $p_j - \tilde{p}_j$ near to 0, we find that for n large enough,

$$\begin{aligned}|\bar{\tau}_1(P, n) - \bar{\tau}_1(\tilde{P}, n)| &\leq c_1 n\varepsilon/\xi + (c_1 n\varepsilon/\xi)n(1+\lambda)\sum_{j=1}^S \tilde{p}_j^2 \\ &\leq (c_1 n\varepsilon/\xi) + (c_1 n\varepsilon/\xi)n(1+\lambda)\max_{j=1,\dots,S} \tilde{p}_j \\ &= c_1 n\varepsilon/\xi + (c_1 n\varepsilon/\xi)n(1+\lambda)(1 + O(c_1\varepsilon/\xi))\max_{j=1,\dots,S} p_j \\ &\leq 3c_1 n\varepsilon.\end{aligned}$$

This estimate combined with (B.3), (B.4) and (22) completes the proof for the first inequality of the proposition. The second inequality simply follows from the first by an application of Markov's inequality.

B.3. Proof of Proposition 4. By a simple application of Bernstein's inequality, we get that

$$\mathbb{P}\big(Q(\boldsymbol{U}) \notin \mathscr{P}'\big) = \mathbb{P}\big(|\textstyle\sum_{j=1}^{S} U_j - 1| > c_1\varepsilon/\xi\big) \leq 2\exp\Big\{ - \frac{1}{2}\frac{c_1^2\varepsilon^2/\xi^2}{\xi S^{-1} + \frac{1}{3}S^{-1}c_1\varepsilon}\Big\}.$$

Then the conclusion follows from simple algebraic manipulations.

B.4. Proof of Proposition 5. By definition, we have that

$$\bar{\tau}_1(Q(\boldsymbol{U}), n) = n\sum_{j=1}^{S} U_j e^{-n(1+\lambda)U_j}.$$

Whence, $\bar{\tau}_1(Q(\boldsymbol{U}))$ is a sum of i.i.d. random variables taking values in $[0, n\xi S^{-1}]$. By Hoedffding's inequality,

$$\mathbb{P}\big(|\bar{\tau}_1(Q(\boldsymbol{U}), n) - \mathbb{E}[\bar{\tau}_1(Q(\boldsymbol{U}), n)]| > n\varepsilon/2\big) \leq 2\exp\Big\{ - \frac{S\varepsilon^2}{2\xi}\Big\}$$

The conclusion follows from simple algebraic manipulations.

B.5. Proof of Proposition 6. We first consider the case where $1+\lambda \leq \log n$. Remark that in that case we have,

$$\frac{n'\xi}{S} = \frac{n(1 + c_1\varepsilon/\xi)(2c_0/e)(1+\lambda)\log(n)}{n(1+\lambda)} \leq \frac{(3c_0/e)\log(n)}{1+\lambda}, \tag{B.5}$$

where the last inequality is true for n large enough. Using [29, Lemma 6], and because $0 \leq U, V \leq \xi S^{-1}$ almost surely, we find that,

$$\begin{aligned} S\mathsf{TV}&\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \\ &\leq \frac{S}{(L+2)!}\Big(\frac{n'\xi}{2S}\Big)^{L+2}\Big(2 + 2^{n'\xi/(2S)-L} + 2^{n'\xi/(2\log(2)S)-L}\Big) \\ &= \frac{2S(1+o(1))}{(L+2)!}\Big(\frac{n'\xi}{2S}\Big)^{L+2}, \end{aligned}$$

where the last line is a consequence of the definition of L and (B.5). Indeed we always have

$$\frac{n'\xi}{2S} < \frac{n'\xi}{2S\log(2)} \leq \frac{3c_0}{2e\log 2}\log(n) < 0.8c_0\log(n) < L,$$

where the last inequality is again also true at least for n large enough, because $L = c_0 A(\lambda, n) > c_0(1+o(1))\log(n)$, for any choice of $c_0 > 0$ because

$a \log a > 0 \Rightarrow a > 1$.
Now, observing that $(L+2)! > L^2 L!$, we obtain the upper bound

$$\begin{aligned} &S\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \\ &\qquad \leq \frac{2S(1+o(1))}{L^2}\Big(\frac{n\xi}{2S}\Big)^2 \frac{1}{L!}\Big(\frac{n\xi}{2S}\Big)^L \Big(\frac{n'}{n}\Big)^{L+2} \\ &\qquad \leq \frac{2eS(1+o(1))}{L^2}\Big(\frac{c_0 \log(n)}{e}\Big)^2 \frac{1}{L!}\Big(\frac{c_0\log(n)}{e}\Big)^L, \end{aligned}$$

where the last line follows because $L+2 \leq c_2\xi$, and hence $(1+c_1\varepsilon/\xi)^{L+2} \leq \exp\{c_1c_2\varepsilon\} \leq e$, by the assumption on ε. Then, by Stirling's formula, whenever $n \to \infty$ (and hence L),

$$S\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \leq \frac{2(1+o(1))}{\sqrt{2\pi c_0}eA(\lambda,n)^{5/2}}\frac{n(1+\lambda)}{\sqrt{\log(n)}}A(\lambda,n)^{-L}.$$

The conclusion then follows for $c_3 > 0$ large enough by the definition of $A(\lambda, n)$, and because when $1+\lambda \leq \log(n)$ it holds $A(\lambda,n) = \omega(1+o(1))$ with ω solution to $\omega\log\omega = c_0^{-1}$, hence c_3 can be chosen to depends only on c_0.

We now consider the case where $1+\lambda > \log(n)$. Under this constraint $\xi = (2c_0/e)\log^2(n)$, and proceeding as for (B.5) we find that $n'\xi/S \leq (3c_0/e)$ as long as n gets large enough. Whence, whenever n is large enough we certainly have $n'\xi/S = o(L)$, and still by [29, Lemma 6], and along similar lines as in the previous paragraph, we get

$$\begin{aligned} &S\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \\ &\qquad \leq \frac{2S(1+o(1))}{(L+2)!}\Big(\frac{n\xi}{2S}\Big)^{L+2}\Big(\frac{n'}{n}\Big)^{L+2} \\ &\qquad \leq \frac{2e(1+o(1))}{(L+2)!} n(1+\lambda)\Big(\frac{c_0\log^2(n)}{e(1+\lambda)}\Big)^{L+2} \\ &\qquad \leq \frac{2c_0^2(1+o(1))}{e\sqrt{2\pi}}\frac{n\log^4(n)}{1+\lambda}\frac{1}{L^{5/2}}\Big(\frac{c_0\log^2(n)}{(1+\lambda)L}\Big)^L \\ &\qquad \leq \frac{2c_0^2(1+o(1))}{e\sqrt{2\pi}}\frac{n\log^3(n)}{L^{5/2}}\Big(\frac{c_0\log(n)}{L}\Big)^L. \end{aligned}$$

but $L \geq 2c_0\log(n)$, so that the previous bound always goes to zero when $n \to \infty$, and hence gets smaller than $1/2$ for n large enough.

B.6. Proof of Proposition 7. We define the function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ such that $\varphi(x) = x\log(x)$. When $1+\lambda \leq \log(n)$, it is clear that $A(\lambda, n)$ converges to the solution of $\varphi(x) = c_0^{-1} = e$, hence $A(\lambda, n) \to e$, which proves the first claim.

For the second claim, let define,

$$\Delta_n := e\frac{\log(1+\lambda) - (1/2)\log\log(n) + \log(c_3)}{\log(n)}.$$

For n large enough such that $\Delta_n > -1$, it is clear than $A(\lambda, n) \geq 0$. Furthermore, by a Taylor expansion of φ near $x = e$, we find that there is a $\bar{x}$ in the line segment between $A(\lambda, n)$ and e,

$$\begin{aligned}\varphi(A(\lambda,n)) &= \varphi(e) + \varphi'(e)(A(\lambda,n) - e) + \frac{\varphi''(\bar{x})}{2}(A(\lambda,n)-e)^2 \\ &\geq \varphi(e) + \varphi'(e)(A(\lambda,n) - e),\end{aligned}$$

because $\varphi''(x) = 1/x > 0$ whenever $x > 0$. Since $\varphi(A(\lambda,n)) - \varphi(e) = \Delta_n$, $\varphi(e) = e$, and $\varphi'(e) = 2$, we deduce that for those n large,

$$0 \leq A(\lambda, n) \leq e + \Delta_n/2.$$

Therefore,

$$\begin{aligned}e^{-1}A(\lambda,n)^2\log(n) &\leq e\log(n) + \Delta_n\log(n) + \frac{\Delta_n^2\log(n)}{4e} \\ &= e\log(n) + e\log\frac{c_3(1+\lambda)}{\sqrt{\log(n)}} + o(1).\end{aligned}$$

This concludes the proof.

APPENDIX C: EXISTENCE OF RANDOM VARIABLES

Here we prove the existence of the random variables U and V which have been used to construct the prior for determining the minimax lower bound in Section 3.1. More precisely, we prove the following theorem.

THEOREM C.1. *Let $S, L \in \mathbb{N}$ and $\xi > 0$ chosen as in Section 3.1. Then there exist two random variables U and V taking values in $[0, \xi S^{-1}]$ such that when n is large enough,*

$$\begin{gathered}\mathbb{E}[U^k] = \mathbb{E}[V^k] \quad \forall k \in \{0, \ldots, L+1\}, \\ \mathbb{E}[U] = \mathbb{E}[V] = S^{-1}, \quad \mathrm{Var}(U) \leq \xi S^{-2}, \quad \mathrm{Var}(V) \leq \xi S^{-2}, \\ \mathbb{E}[Ue^{-n(1+\lambda)U}] \geq \mathbb{E}[Ve^{-n(1+\lambda)V}] + S^{-1}K\min\left\{1, \sqrt{\xi/L^2}\exp(-L^2/\xi)\right\}.\end{gathered}$$

The proof of Theorem C.1 follows the guidelines used in the papers Wu and Yang [29, 28], relating the problem of the existence of the random variables to the problem of finding the best polynomial approximation to some function.

For $a, b \in \mathbb{R}$, we let $\mathsf{C}[a, b]$ denote the space of continuous functions on $[a, b]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{P}_L[a, b] \subset \mathsf{C}[a, b]$ denote the space of polynomials of degree no more than L on $[a, b]$. For any $f \in \mathsf{C}[a, b]$, the best polynomial (of degree at most L) approximation to f is defined as

$$E_L(f, [a, b]) := \inf\{\sup\{|f(x) - q(x)| : \ x \in [a, b]\} : \ q \in \mathsf{P}_L[a, b]\}.$$

For the sake of simplicity, we define $B := n(1 + \lambda)\xi/(2S)$. Remark that $B \asymp \xi/2$, but is not necessarily equal to it because S is integer. We define $g : [\xi^{-1}, 1] \to \mathbb{R}_+$ such that $g(x) := \exp\{-2Bx\}$. It is a classical result that for any $L \in \mathbb{N}$ we can find random variables X and Y taking values in $[\xi^{-1}, 1]$ and such that

$$\begin{aligned}\mathbb{E}[X^k] &= \mathbb{E}[Y^k], \qquad k = 0, \dots, L,\\ \mathbb{E}[g(X)] &= \mathbb{E}[g(Y)] + E_L(g, [\xi^{-1}, 1]).\end{aligned}$$

The proof of the existence of such random variables can be found for instance in Wu and Yang [28, 29] for a constructive argument, or for instance in Lepski et al. [11] using the Hahn-Banach theorem and a duality argument.

We now assume that we have at our disposal the random variables X and Y of the previous paragraph, and we write P_X and P_Y their distributions. The construction of the random variables U and V is done using the trick introduced in Wu and Yang [28, Lemma 4]. Namely, we let U and V having respective distributions on $[0, \xi S^{-1}]$

$$\begin{aligned}P_U(\mathrm{d}x) &:= \big(1 - \mathbb{E}[(\xi X)^{-1}]\big)\delta_0 + (Sx)^{-1} P_{\xi X/S}(\mathrm{d}x),\\ P_V(\mathrm{d}x) &:= \big(1 - \mathbb{E}[(\xi Y)^{-1}]\big)\delta_0 + (Sx)^{-1} P_{\xi Y/S}(\mathrm{d}x).\end{aligned}$$

Because $X, Y \geq \xi^{-1}$ almost-surely, then $\mathbb{E}[(\xi X)^{-1}] \leq 1$ and $\mathbb{E}[(\xi Y)^{-1})] \leq 1$. Indeed from Wu and Yang [28, Lemma 4], P_U and P_V are proper probability distributions on $[0, \xi S^{-1}]$ satisfying

$$\begin{aligned}&\mathbb{E}[U] = \mathbb{E}[V] = 1/S, \qquad \mathbb{E}[U^k] = \mathbb{E}[V^k], \qquad k = 0, \dots, L + 1,\\ &\mathbb{E}[U \exp\{-n(1 + \lambda)U\}] = \mathbb{E}[V \exp\{-n(1 + \lambda)V\}] + S^{-1} E_L(g, [\xi^{-1}, 1]).\end{aligned}$$

Furthermore, it is clear that,

$$\mathbb{E}[U^2] = \frac{1}{S} \int x \, P_{\xi x/S}(\mathrm{d}x) = \frac{\xi \mathbb{E}[X]}{S^2} \leq \frac{\xi}{S^2}.$$

Hence $\mathrm{Var}(U) \le \xi/S^2$. It is obvious that we also have $\mathrm{Var}(V) \le \xi/S^2$. Thus, the proof of the theorem is finished by obtaining a lower bound on the best polynomial approximation $E_L(g, [\xi^{-1}, 1])$, which is done in the next section (in particular see Theorem D.1 and the (D.2) just after).

APPENDIX D: APPROXIMATION THEORY

D.1. Statement of the main result. We wish to find the best polynomial approximation (see Section C for definition) to the function $g : [\xi^{-1}, 1] \to \mathbb{R}_+$ such that $g(x) := \exp\{-2Bx\}$ on $[\xi^{-1}, 1]$, with ξ defined in (24) and B satisfying

$$\frac{\xi/2}{1 + \frac{1}{n(1+\lambda)}} \le B \le \xi/2. \tag{D.1}$$

The whole section will be dedicated to the proof of the following theorem.

THEOREM D.1. *For every $\zeta > 0$, there exists a constant $K > 0$ such that as $n \to \infty$,*

$$E_L(g, [\xi^{-1}, 1]) \ge K(1 + o(1)) \cdot \begin{cases} 1 & \text{if } L \le \sqrt{\xi/2}, \\ \frac{\sqrt{\xi}\exp\{-L^2/\xi\}}{L(1+(2L/\xi)^2)^{1/4}} & \text{if } \sqrt{\xi/2} < L < \zeta\xi. \end{cases}$$

We deduce from the previous theorem that there exists a universal constant $K > 0$ such that for n large enough,

$$E_L(g, [\xi^{-1}, 1]) \ge K \min\{1, \sqrt{\xi/L^2}\exp(-L^2/\xi)\}. \tag{D.2}$$

D.2. Proof of Theorem D.1. Let $\sigma : [-1, 1] \to [\xi^{-1}, 1]$ be such that $\sigma(x) := (1 - \xi^{-1})(x + 1)/2 + \xi^{-1}$. Notice that σ is bijective. By translating and rescaling, we claim that $E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1])$. To see that this is true, remark that for all $p \in \mathsf{P}_L[-1, 1]$ we have $\|g \circ \sigma - p\|_\infty = \|g - p \circ \sigma^{-1}\|_\infty \ge E_L(g, [\xi^{-1}, 1])$. This shows that $E_L(g \circ \sigma, [-1, 1]) \ge E_L(g, [\xi^{-1}, 1])$. The same steps using σ^{-1} show that $E_L(g \circ \sigma, [-1, 1]) \le E_L(g, [\xi^{-1}, 1])$. Hence $E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1])$.

For the sake of simplicity, we let $C := B(1 - \xi^{-1})$ and $\gamma : [-1, 1] \to \mathbb{R}_+$ is defined by $\gamma(x) = \exp\{-C(x + 1)\}$. From the discussion in the previous paragraph, we have indeed reduced the problem to finding $E_L(\gamma, [-1, 1])$. This is because

$$E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1]) = \exp\{-2B\xi^{-1}\}E_L(\gamma, [-1, 1]). \tag{D.3}$$

To find a lower bound on $E_L(\gamma, [-1,1])$, we will exploit the well-known relationship between uniform approximation on the interval by polynomials and uniform approximation of periodic even functions by trigonometric polynomials. We write $\mathsf{CE}[-1,1]$ the space of continuous and even functions on $[-1,1]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{TP}_L[-1,1]$ denote the set of even trigonometric polynomials of degree at most L, *i.e.* $\mathsf{TP}_L[-1,1]$ is

$$\Big\{T \in \mathsf{CE}[-1,1] :\ T(x) = \textstyle\sum_{k=0}^{L} a_k \cos(\pi k x),\ a_k \in \mathbb{R},\ x \in [-1,1]\Big\}.$$

We furthermore define the periodization operator $P : \mathsf{C}[-1,1] \to \mathsf{CE}[-1,1]$ such that $Pf(\theta) = f(\cos(\pi\theta))$ for all $f \in \mathsf{C}[-1,1]$ and all $\theta \in [-1,1]$. Then, it is well-known (see for instance the Theorem 14.8.1 in [6]) that

$$E_L(\gamma, [-1,1]) = \inf\{\|P\gamma - T\|_\infty :\ T \in \mathsf{TP}_L[-1,1]\}. \tag{D.4}$$

We will now bound the r.h.s. of (D.4) by a technique inspired from Newman and Rivlin [16], which surprisingly work as well for our setting. For any $K \in \mathbb{N}$, we define the trigonometric polynomial $T_K : [-1,1] \to \mathbb{C}$ such that

$$T_K(\theta) := e^{i\pi(L+1)\theta}\Big\{\sum_{k=0}^{K-1} e^{i2\pi k\theta}\Big\}^2.$$

Then, by orthogonality of the trigonometric polynomials, we have that

$$\int_{-1}^{-1} |T_K(\theta)|\,\mathrm{d}\theta = \sum_{j=0}^{K-1}\sum_{k=0}^{K-1}\int_{-1}^{1} e^{i2\pi(j-k)\theta}\,\mathrm{d}\theta = K. \tag{D.5}$$

By definition, for every $\varepsilon > 0$ we can find a $Q \in \mathsf{TP}_L[-1,1]$ such that $\|P\gamma - Q\|_\infty \le E_L(\gamma,[-1,1]) + \varepsilon$. Choose such Q, and remark that (D.5) implies,

$$\begin{aligned}\Big|\int_{-1}^{1}(P\gamma(\theta) - Q(\theta))T_K(\theta)\,\mathrm{d}\theta\Big| &\le \|P\gamma - Q\|_\infty \int_{-1}^{1} |T_K(\theta)|\,\mathrm{d}\theta\\ &\le K\{E_L(\gamma,[-1,1]) + \varepsilon\}.\end{aligned}$$

On the other hand remark that Q is a trigonometric polynomial of degree at most L, while T_K is a trigonometric polynomial of degree strictly greater than L. Therefore Q is orthogonal to T_K. Moreover, the last display is true for all $\varepsilon > 0$ and for all $K \in \mathbb{N}$, thus it must be the case that

$$E_L(\gamma,[-1,1]) \ge \max_{K\in\mathbb{N}} \frac{1}{K}\Big|\int_{-1}^{1} P\gamma(\theta)T_K(\theta)\,\mathrm{d}\theta\Big|. \tag{D.6}$$

Interestingly, we can compute the previous integral. Namely,

$$\begin{aligned}\int_{-1}^{1} P\gamma(\theta)T_K(\theta)\,\mathrm{d}\theta &= \sum_{j=0}^{K-1}\sum_{k=0}^{K-1}\int_{-1}^{1}\gamma(\cos(\pi\theta))e^{i\pi\theta(L+1+2j+2k)}\,\mathrm{d}\theta \\ &= \sum_{j=0}^{K-1}\sum_{k=0}^{K-1}\int_{0}^{1}\gamma(\cos(\pi\theta))\cos(\pi\theta(L+1+2j+2k))\,\mathrm{d}\theta\end{aligned}$$

The integrals involved in the last display can be expressed in terms of the modified Bessel function (see [1, pg. 248]) denoted here as $I_\nu(z)$, which equals

$$I_k(z) = \frac{1}{\pi}\int_0^\pi e^{z\cos(t)}\cos(\nu t)\mathrm{d}t$$

whenever $\nu = k \in \mathbb{N}$ thanks to [1, formula 10.32.3]. More precisely, from the above considerations and the fact that the modified Bessel functions are non–negative, we deduce that

$$\Big|\int_{-1}^{1} P\gamma(\theta)T_K(\theta)\,\mathrm{d}\theta\Big| = \sum_{j=0}^{K-1}\sum_{k=0}^{K-1} e^{-C}I_{L+1+2j+2k}(C).$$

Soni [25] proved that $I_{k+1}(z) \leq I_k(z)$ for all $k \in \mathbb{N}$ and all $z > 0$. Hence, we obtain from the last display and (D.6) the bound

$$E_L(\gamma, [-1,1]) \geq \max_{K\in\mathbb{N}} Ke^{-C}I_{L+4K}(C). \tag{D.7}$$

In the next lemma, We obtain a bound on the modified Bessel function $z \mapsto I_k(z)$ which remains tighter than the classical bound derived in Luke [12] when $z \geq k$. The proof of the lemma is to be found in Section D.3.

LEMMA D.1. *Assume $k \in \mathbb{N}$ and assume that $z > 8\sqrt{1+(k/z)^2}$. Then,*

$$e^{-z}I_k(z) > \frac{\exp\{-k^2/(2z)\}}{2e^4(1+(k/z)^2)^{1/4}\sqrt{z}}.$$

$$K_* := \begin{cases}\alpha\sqrt{C} & \text{if } L < \sqrt{C},\\ \beta C/L & \text{if } L \geq \sqrt{C}.\end{cases}$$

In view of (D.7), it is clear that $E_L(\gamma,[-1,1]) \geq K_*e^{-C}I_{L+4K_*}(C)$. Consider now the case where $L < \sqrt{C}$, then

$$0 \leq \frac{L+4K_*}{C} = \frac{L+\alpha\sqrt{C}}{C} < \frac{\alpha+1}{\sqrt{C}}.$$

Thus, $(L+4K_*)/C \to 0$ as $C \to \infty$, and this implies that $C > 8\sqrt{1 + (L + 4K_*)^2/C^2}$ when C gets large enough. We then obtain from Lemma D.1 that in this case, as $C \to \infty$,

$$\begin{aligned} E_L(\gamma, [-1, 1]) &> \frac{\alpha\sqrt{C}(1 + o(1)) \exp\{-(L + 4K_*)^2/(2C)\}}{2e^2\sqrt{C}} \\ &> \frac{\alpha(1 + o(1)) \exp\{-(\alpha + 1)^2/2\}}{2e^2} \gtrsim 1. \end{aligned}$$

We now consider the case $L \geq \sqrt{C}$. In this case, we have,

$$0 \leq \frac{L + 4K_*}{C} = \frac{L + \beta C/L}{C} = \frac{L}{C} + \frac{\beta}{L} \leq \frac{L}{C} + \frac{\beta}{\sqrt{C}}.$$

Because by assumption there is a constant $\zeta > 0$ such that $L \leq \zeta C$, then $(L+4K_*)/C \leq \zeta + o(1)$ as $C \to \infty$, and thus we have $C > 8\sqrt{1 + (L + 4K_*)^2/C^2}$ when C is large enough. Then, we can apply Lemma D.1 to find that as $C \to \infty$, because $K_*^2/C = \beta^2 C/L^2 \lesssim 1$ and $K_* L/C = \beta \lesssim 1$,

$$\begin{aligned} E_L(\gamma, [-1, 1]) &> \frac{(\beta C/L)(1 + o(1)) \exp\{-(L + 4K_*)^2/(2C)\}}{2e^2\sqrt{C}(1 + (L/C)^2)^{1/4}} \\ &= \frac{\beta\sqrt{C}(1 + o(1)) \exp\{-L^2/(2C) - 8K_*^2/C - 4K_* L/C\}}{2e^2 L(1 + (L/C)^2)^{1/4}} \\ &\gtrsim \frac{\sqrt{C}(1 + o(1)) \exp\{-L^2/(2C)\}}{L(1 + (L/C)^2)^{1/4}}. \end{aligned}$$

The conclusion then follows from (D.3), by remarking that $2B\xi^{-1} = 1 + o(1)$ and $C = B(1 - \xi^{-1}) = (\xi/2)(1 - \xi^{-1})(1 + O(1/n))$ as $n \to \infty$, and thus $L^2/(2C) \leq L^2/\xi + C$ for some $C > 0$ by definitions of L and ξ, and $L/C = 2L(1 + o(1))/\xi$.

D.3. Proof of Lemma D.1. The proof relies on the well known series representation of the modified Bessel function (see [1, formula 10.25.2]), namely we have whenever $k \in \mathbb{N}$,

$$I_k(z) = \sum_{p=0}^{\infty} \frac{1}{p!(p + k)!} \left(\frac{z}{2}\right)^{2p+k}. \tag{D.8}$$

Conveniently, all the terms in the summation are non-negative, which we will exploit to get our lower bound.

By Stirling's formula, when $k \geq 1$, for any $p \geq 0$

$$(p + k)! \leq e\sqrt{(p + k)} \exp\{-(p + k) + (p + k) \log(p + k)\},$$

and for any $p \geq 1$,

$$p! \leq e\sqrt{p}\exp\{-p + p\log p\}.$$

For convenience, let define the functions $\phi_{z,k} : \mathbb{R}_+ \to \mathbb{R}_+$, such that for any $x, z \in \mathbb{R}_+$ and any $k \in \mathbb{N}$,

$$\phi_{z,k}(x) := -z + 2x + k - x\log x - (x+k)\log(x+k) + (2x+k)\log(z/2).$$

Hence, because each term in the series expansion of (D.8) is non-negative, we get the estimate,

$$e^{-z}I_k(z) \geq e^{-z}\sum_{p\geq 1}\frac{1}{p!(p+k)!}\Big(\frac{z}{2}\Big)^{2p+k} \geq \frac{1}{e^2}\sum_{p\geq 1}\frac{\exp\{\phi_{z,k}(p)\}}{\sqrt{p(p+k)}}. \tag{D.9}$$

Notice that,

$$\phi'_{z,k}(x) = -\log(x) - \log(x+k) + 2\log(z/2), \qquad \phi''_{z,k}(x) = -\frac{1}{x} - \frac{1}{x+k}.$$

Thus, $\phi_{z,k}$ admits a unique non-negative extremum at x_0 solution to $x_0(x_0 + k) = z^2/4$, that is,

$$x_0 = \frac{-k + \sqrt{k^2 + z^2}}{2}, \text{ and, } \quad \phi''_{z,k}(x_0) = -\frac{4}{z}\sqrt{1 + (k/z)^2} < 0.$$

Henceforth x_0 is indeed the unique maximum of the function $\phi_{z,k}$ on $\mathbb{R}_+$. We let p_0 smallest integer larger than x_0. Then $p_0 \geq 1$ and we have, by Taylor expansion that for any $p \geq p_0$ there is a $\bar{p} \in (x_0, p)$

$$\begin{aligned}\phi_{z,k}(p) &= \phi_{z,k}(x_0) + \phi'_{z,k}(x_0)(p - x_0) + \frac{1}{2}\phi''_{z,k}(\bar{p})(p - x_0)^2 \\ &= \phi_{z,k}(x_0) + \frac{1}{2}\phi''_{z,k}(\bar{p})(p - x_0)^2.\end{aligned}$$

Remark that, because $\bar{p} \geq x_0$,

$$\phi''_{z,k}(\bar{p}) = -\frac{1}{\bar{p}} - \frac{1}{\bar{p} + k} \geq -\frac{1}{x_0} - \frac{1}{x_0 + k} = -\frac{4}{z}\sqrt{1 + (k/z)^2}.$$

Then, for any $p \geq p_0$,

$$\begin{aligned}\phi_{z,k}(p_0) &\geq \phi_{z,k}(x_0) + \frac{1}{2}\phi''_{z,k}(x_0)(p_0 - x_0)^2 \\ &= \phi_{z,k}(x_0) - \frac{2\sqrt{1 + (k/z)^2}}{b}(p - x_0)^2.\end{aligned}$$

Therefore,

$$e^{-z} I_k(z) \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^2} \sum_{p \geq p_0} \frac{\exp\{\phi''_{z,k}(x_0)(p - x_0)^2/2\}}{\sqrt{p(p+k)}}.$$

Let p_1 be the largest integer such that $-\phi''_{z,k}(x_0)(p_1 - x_0)^2 \leq 2$. Remark that whenever $z > 2(1 + (k/z)^2)^{1/2}$, we have $p_1 \geq x_0 + 1$, which is always the case in the conditions of the lemma. Because the summand is the previous is monotonically decreasing for $p \geq p_0$, we get the bound,

$$e^{-z} I_k(z) \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4} \frac{(p_1 - p_0)}{\sqrt{p_1(p_1 + k)}} \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4} \frac{(p_1 - x_0) - 1}{\sqrt{p_1(p_1 + k)}}.$$

But, by the definition of p_1, we have that,

$$p_1 + 1 - x_0 > \sqrt{\frac{2}{-\phi''_{z,k}(x_0)}}.$$

Therefore, whenever $z > 8(1 + (k/z)^2)^{1/2}$, by the definition of $\phi''_{z,k}(x_0)$,

$$\begin{aligned} e^{-z} I_k(z) &\geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4 \sqrt{-\phi''_{z,k}(x_0) p_1(p_1 + k)}} \left\{ \sqrt{2} - 2\sqrt{-\phi''_{z,k}(x_0)} \right\} \\ &\geq \frac{\sqrt{2} \exp\{\phi_{z,k}(x_0)\}}{2e^4 \sqrt{-\phi''_{z,k}(x_0) p_1(p_1 + k)}}. \end{aligned}$$

Also,

$$\begin{aligned} p_1(p_1 + k) &= x_0(x_0 + k) + (p_1^2 - x_0^2) + (p_1 - x_0)k \\ &= x_0(x_0 + k) + (p_1 - x_0)(p_1 + x_0 + k) \\ &= x_0(x_0 + k) + (p_1 - x_0)^2 + (p_1 - x_0)(2x_0 + k). \end{aligned}$$

But we have that $x_0(x_0 + k) = z^2/4$, $(p_1 - x_0)^2 \leq -2/\phi''_{z,k}(x_0)$, and $2x_0 + k = z\sqrt{1 + (k/z)^2}$. Thus,

$$\begin{aligned} p_1(p_1 + k) &\leq \frac{z^2}{4} + \frac{2}{-\phi''_{z,k}(x_0)} + \sqrt{\frac{2(1 + (k/z)^2)}{-\phi''_{z,k}(x_0)}} z \\ &= \frac{z^2}{4} + \frac{z}{2\sqrt{1 + (k/z)^2}} + \frac{z^{3/2}}{\sqrt{2}} [1 + (k/z)^2]^{1/4} \end{aligned}$$

$$= \frac{z^2}{4}\Big\{1 + \frac{z^{-1/2}[1+(k/z)^2]^{1/4}}{\sqrt{2}} + \frac{z^{-1}}{2\sqrt{1+(k/z)^2}}\Big\}.$$

Therefore, whenever $z > 8(1+(k/z)^2)^{1/2}$,

$$p_1(p_1+k) \le \frac{z^2}{4}\Big\{1 + \frac{1}{4} + \frac{1}{16}\Big\} \le \frac{21}{64}z^2 < \frac{z^2}{2}.$$

Hence,

$$e^{-z}I_k(z) > \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4\sqrt{-\phi''_{z,k}(x_0)z}} = \frac{\exp\{\phi_{z,k}(x_0)\}}{2e^4(1+(k/z)^2)^{1/4}\sqrt{z}}.$$

The remainder of the proof is now dedicated to deriving a lower bound on $\phi_{z,k}(x_0)$. After some algebra, we find that

$$\begin{aligned}\phi_{z,k}(x_0) = {} & -z + z\sqrt{1+(k/z)^2} \\ & - (z/2)\{-(k/z) + \sqrt{1+(k/z)^2}\}\log\{-(k/z) + \sqrt{1+(k/z)^2}\} \\ & - (z/2)\{(k/z) + \sqrt{1+(k/z)^2}\}\log\{(k/z) + \sqrt{1+(k/z)^2}\}.\end{aligned}$$

Now we define the function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ such that

$$\begin{aligned}\varphi(x) := -1 + \sqrt{1+x^2} - \frac{1}{2}(-x+\sqrt{1+x^2})\log(-x+\sqrt{1+x^2}) \\ - \frac{1}{2}(x+\sqrt{1+x^2})\log(x+\sqrt{1+x^2}).\end{aligned}$$

Notice that $\phi_{z,k}(x_0) = z\varphi(k/z)$. Also,

$$\varphi'(x) = \frac{(-x+\sqrt{1+x^2})\log(-x+\sqrt{1+x^2})}{2\sqrt{1+x^2}} - \frac{(x+\sqrt{1+x^2})\log(x+\sqrt{1+x^2})}{2\sqrt{1+x^2}},$$

$$\varphi''(x) = -\frac{1}{(1+x^2)^{1/2}}, \qquad \varphi'''(x) = \frac{x}{(1+x^2)^{3/2}}.$$

By a Taylor expansion of φ near 0, we find that there is a $y \in (0,x)$ such that

$$\varphi(x) = \varphi(0) + \varphi'(0)x + \frac{1}{2}\varphi''(0)x^2 + \frac{1}{6}\varphi'''(y)x^3 \ge -\frac{x^2}{2},$$

because $\varphi(0) = \varphi'(0) = 0$ and $\varphi'''(y) \ge 0$ for all $y \ge 0$ by the computations above. This gives the proof for the lower bound on $\phi_{z,k}(x_0)$ as well, concluding the proof.

REFERENCES

[1] Olver, F.W.J., Lozier, D.W., Boisvert, R.F. and Clark, C.W. (2010). *NIST handbook of mathematical functions*, Cambridge University Press.

[2] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *J. Amer. Statist. Assoc.* **85**, 38–45.

[3] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.* **88**, 364–373.

[4] Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Statist.* **9**, 525–546.

[5] Carota, C., Filippone, M. and Polettini, S. (2018). Assessing Bayesian nonparametric log-linear models: an application to disclosure risk estimation. *Preprint: arXiv:1801.05244*

[6] Davidson, K. and Donsig, A. (2009). Real analysis and applications: theory in practice. *Springer Science & Business Media.*

[7] Efron, B. and Morris, C (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117–130.

[8] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

[9] Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1**, 209–230.

[10] Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.

[11] Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory and Related Fields* **113**, 221–253.

[12] Luke, Y.L. (1972). Inequalities for generalized hypergeometric functions. *J. Approximation Theory*, **5**, 41–65.

[13] Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394.

[14] Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079.

[15] Maritz, J.S. and Lwin, T. (1989). *Empirical Bayes methods with applications*, Chapman and Hall.

[16] Newman, D.J. and Rivlin, T.J. (1976). Approximation of monomials by lower degree polynomials. *Aequationes Math.* **14**, 451–455.

[17] Orlitsky, A., Suresh, A.T. and Wu, Y. (2017). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113**, 13283–13288.

[18] Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100**, 1103–1112.

[19] Rinott, Y. and Shlomo, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, Springer, Berlin.

[20] Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.*,**1**, 157–163.

[21] Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Statist.* **14**, 373–383.

[22] Skinner, C.J. and Elliot, M.J. (2002). A measure of disclosure risk for microdata.

J. Roy. Statist. Soc. B **64**, 855–867.

[23] Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *J. Off. Stat.* **10**, 31–51.

[24] Skinner, and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103**, 989–1001.

[25] Soni, R.P. (1965). On an inequality for modified Bessel functions. *J. Math. and Phys.* **44**, 1–4.

[26] Valiant, G. and Valiant, P. (2013). Estimating the unseen: improved estimators for entropy and other properties. *Adv. Neural Inform. Process. Syst. 26*

[27] Willenborg, L. and de Waal, T. (2001) *Elements of statistical disclosure control.* Springer, New York.

[28] Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62**, 3702–3720.

[29] Wu, Y. and Yang, P. (2015). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.*, to appear.

Department of Economics, Management and Statistics
Milano, 20126
Italy
E-mail: federico.camerlenghi@unimib.it

Department of Economics and Statistics
Torino, 10134
Italy
E-mail: stefano.favaro@unito.it

Department of Statistical Sciences, Sidney Smith Hall
Toronto, M5S 3G3
Canada
E-mail: znaulet@utstat.toronto.edu

Department of Statistics
Oxford, OX1 3LB,
United Kingdom
E-mail: francesca.panero@stats.ox.ac.uk